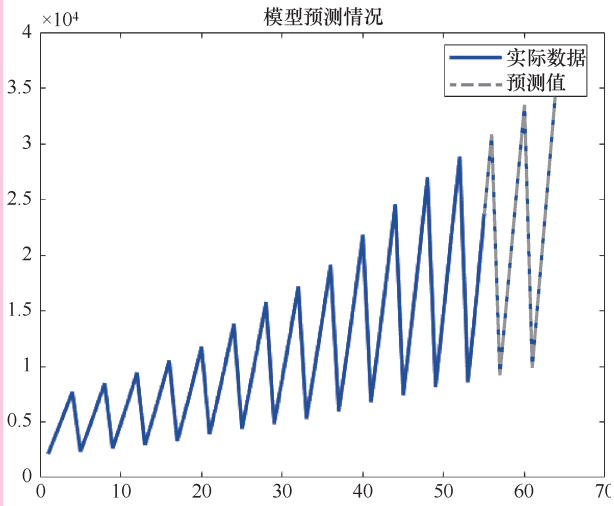


图 10-4　模型预测情况图

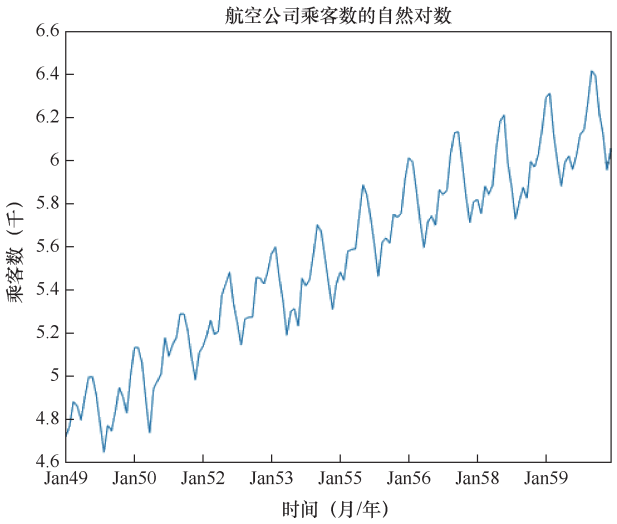


图 11-1　乘客数（自然对数）的时间序列趋势图

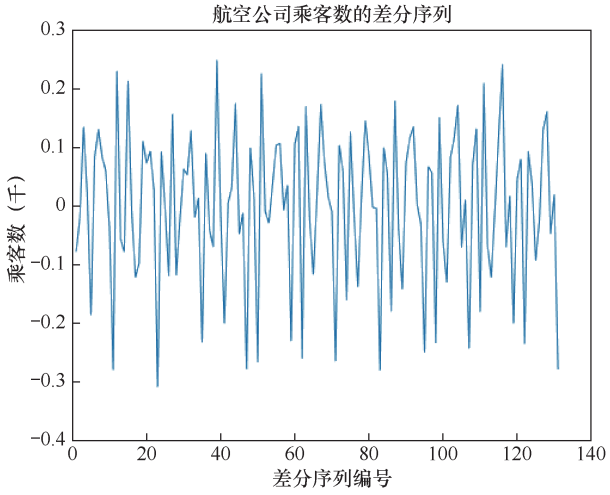


图 11-2　乘客序列差分序列图

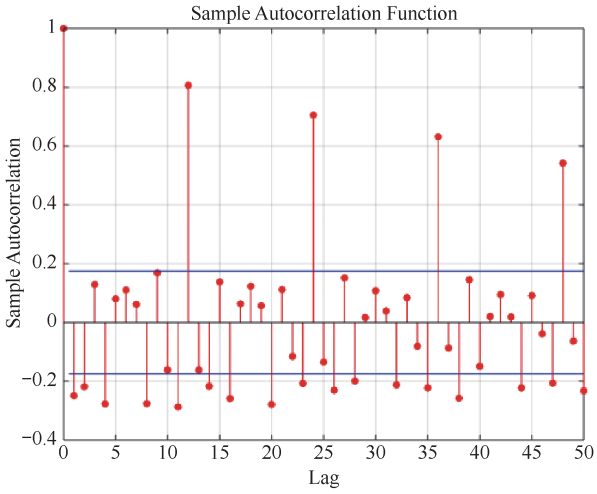


图 11-3　样本的自相关函数图

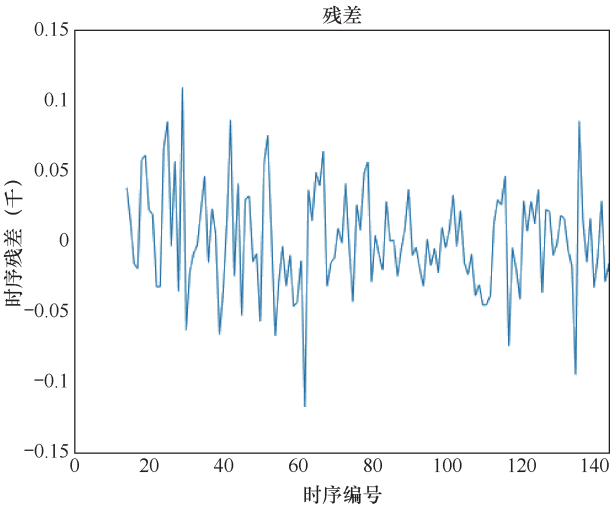


图 11-4 乘客序列的残差图

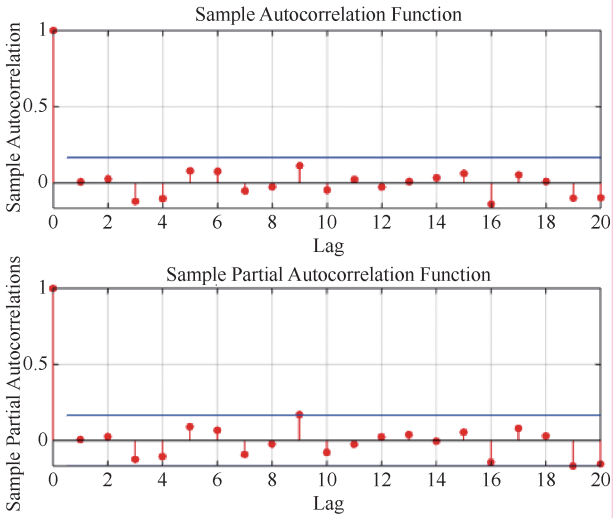


图 11-6 乘客序列的残差分布图

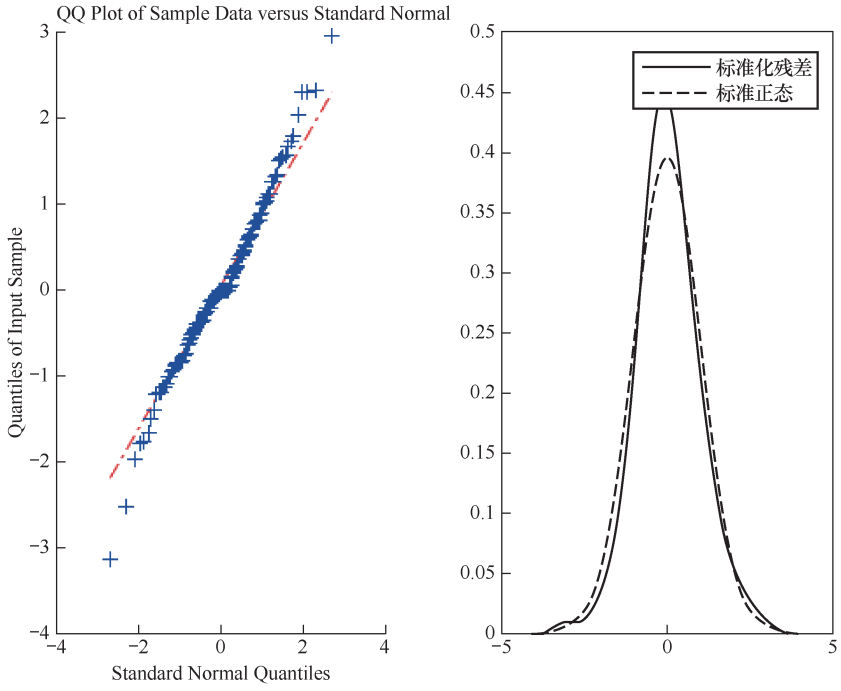


图 11-5 乘客序列的残差分布图

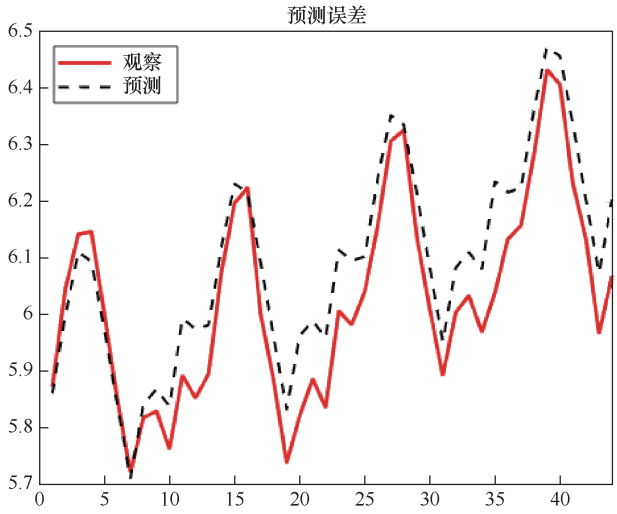


图 11-7　预测与实际数值的比较图

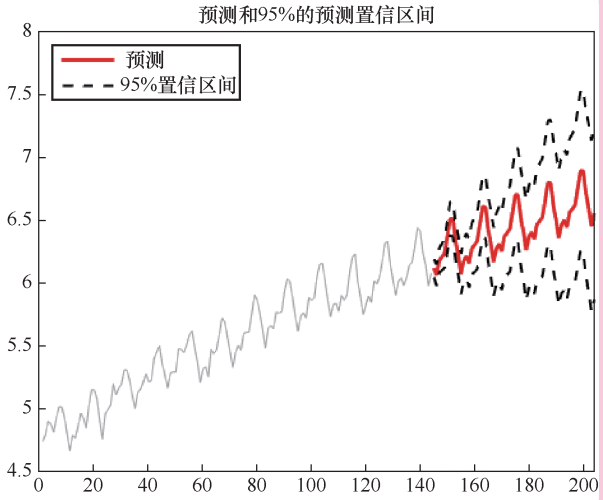


图 11-8　预测的置信区间图

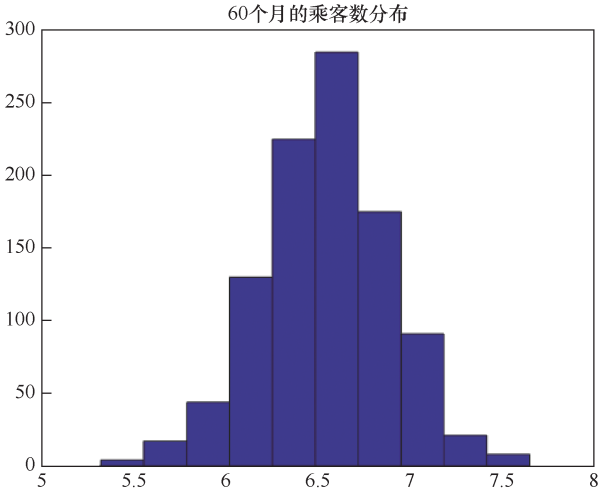


图 11-10　预测结果的分布图

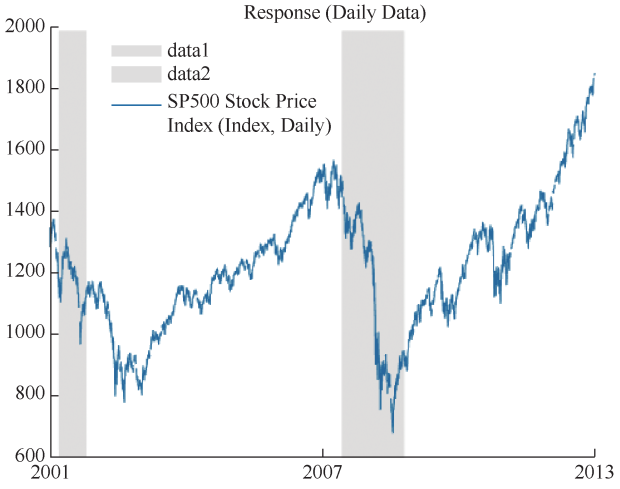


图 12-1　标普 500 走势图

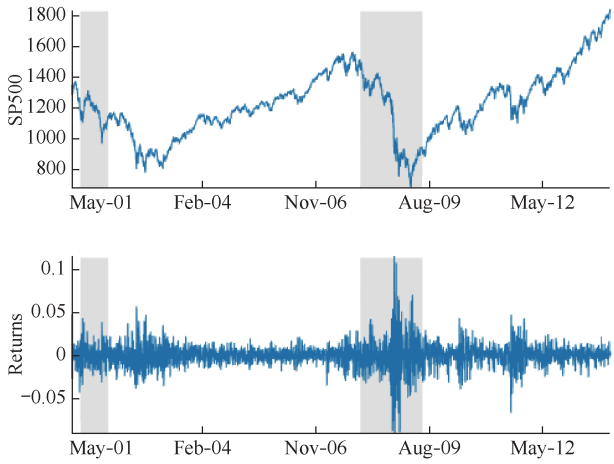


图 12-2　标普 500 收益序列图

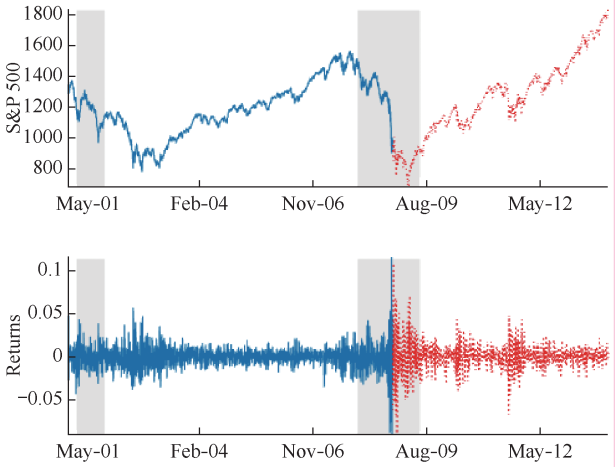


图 12-3　训练集和测试集的划分
（右侧虚线部分为测试集）

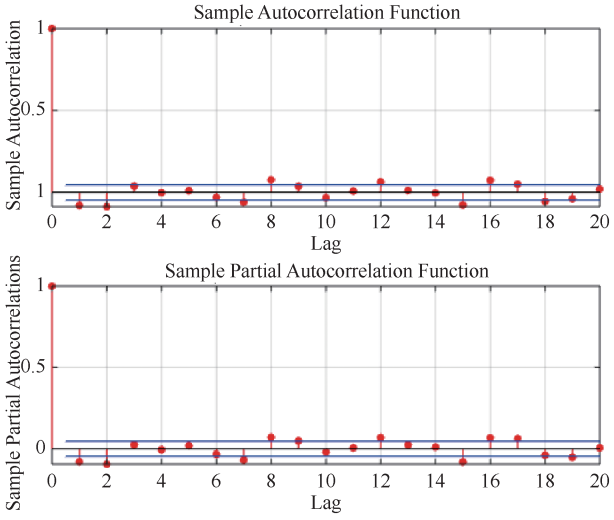


图 12-4　序列的 ACF 和 PACF 图

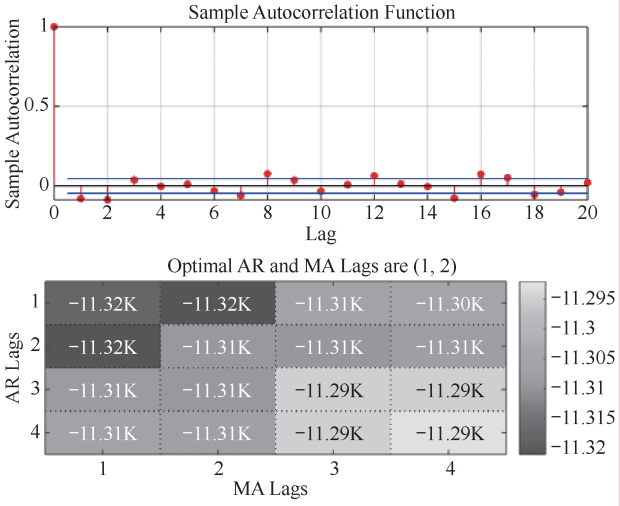


图 12-5　MA 与 ARL 参数扫描结果

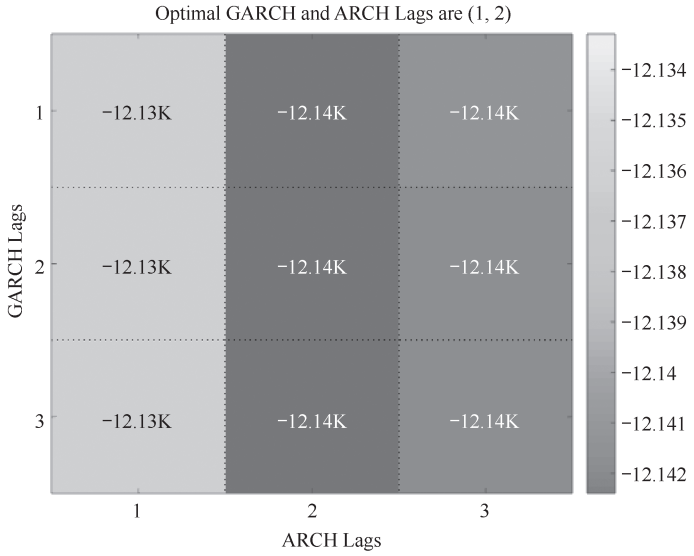


图 12-6　ARCH Lags 与 Garch Lags 参数扫描结果

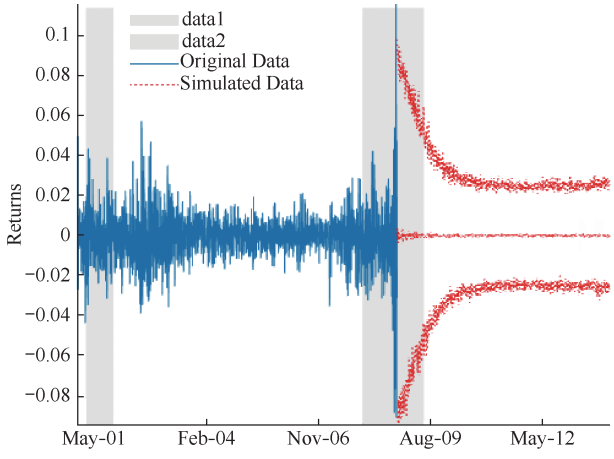


图 12-7　收益的仿真结果

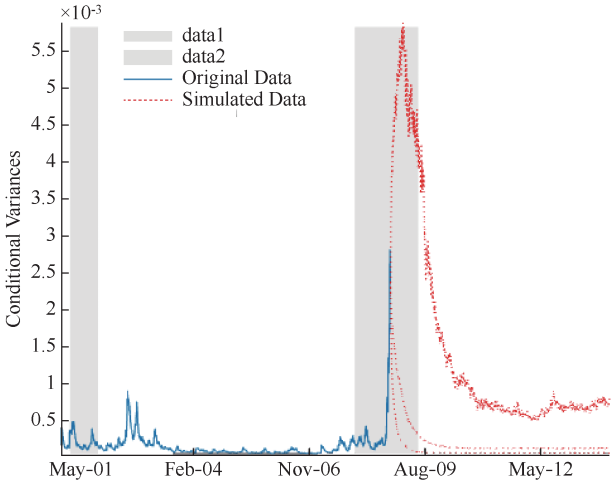


图 12-8　条件方差仿真结果

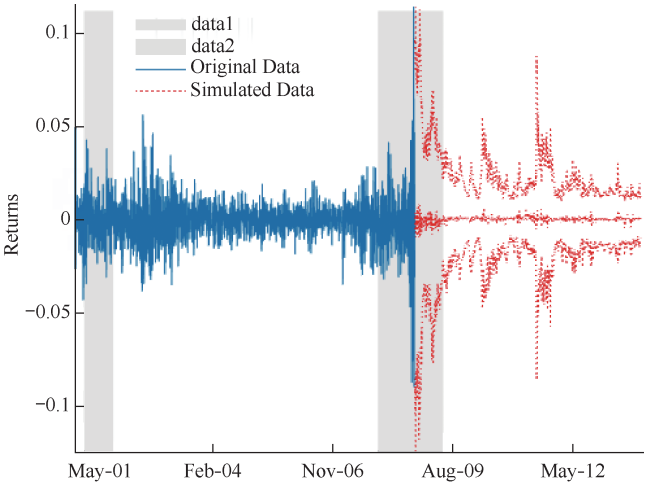


图 12-9　蒙特卡罗仿真结果

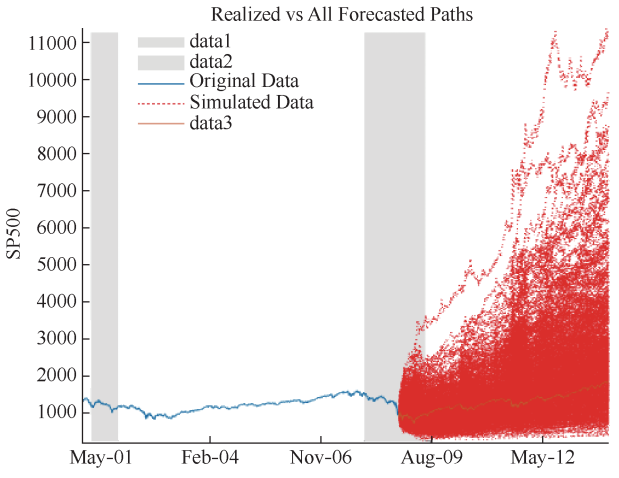


图 12-10　预测轨迹分布图

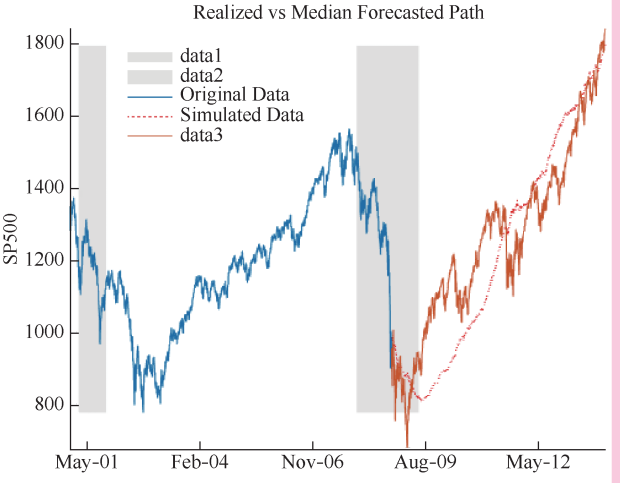


图 12-11　中值价格路径预测结果

大数据金融丛书

Matlab

MATLAB

时间序列方法与实践

江渝　李幸　卓金武　编著

電子工業出版社
Publishing House of Electronics Industry
北京・BEIJING

内 容 简 介

时间序列在金融、经济领域被广泛应用，时间序列的应用一般要依赖各种工具的辅助。MATLAB 的计量经济学工具箱集成了绝大多数的时间序列方法，为时间序列的广泛应用提供了强大的工具支撑。本书将系统介绍时间序列的理论方法与这些方法的 MATLAB 实现过程和实例。

全书内容分三个部分。第一部分为第 1～2 章，主要介绍时间序列的概况和基本概念；第二部分为第 3～10 章，是本书的重点，依次介绍了 AR、MA、ARMA、ARIMA 模型，时间序列平稳性检验，趋势与季节性时间序列建模，ARCH、GARCH 模型和多元时间序列建模的理论方法及这些方法的 MATLAB 实现过程；第三部分为第 11～12 章，介绍了两个时间序列的综合应用实例，通过实例，诠释了概念、理论的实际应用，并给出了全部的 MATLAB 实现代码。

本书适合作为金融、经济、应用数学、统计、大数据等专业的学生和老师的教材或参考用书，也可以作为时间序列领域的科研人员、学者、工程技术人员的参考用书。

图书在版编目（CIP）数据

MATLAB 时间序列方法与实践/江渝，李幸，卓金武编著. —北京：电子工业出版社，2019.4
（大数据金融丛书）
ISBN 978-7-121-36053-4

Ⅰ. ①M… Ⅱ. ①江… ②李… ③卓… Ⅲ. ①Matlab 软件－应用－时间序列分析
Ⅳ. ①O211.61-39

中国版本图书馆 CIP 数据核字（2019）第 034110 号

策划编辑：李　冰
责任编辑：李　冰
文字编辑 1：李　冰　文字编辑 2：冷春雨
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：13.5　字数：233 千字　彩插：3
版　　次：2019 年 4 月第 1 版
印　　次：2019 年 4 月第 1 次印刷
定　　价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：libing@phei.com.cn。

除了你的才华，其他一切都不重要！

近年来，互联网和人工智能技术飞速发展，推动传统金融大踏步前进，尤其在量化投资、互联网金融、移动计算等领域，用一日千里来形容也不为过。2015 年年初，李克强总理在政府工作报告中提出要制定“互联网+”行动计划，推动移动互联网、云计算、大数据等与各行业的融合发展。2015 年 9 月，国务院又印发了《促进大数据发展行动纲要》，纲要提出“推动产业创新发展，培育数据应用新业态，积极推动大数据与其他行业的融合，大力培育互联网金融、数据服务、数据处理分析等新业态”。可见，大数据金融将会成为未来十年最闪亮的领域之一。2012 年年初，中国量化投资学会联合电子工业出版社，共同策划出版了“量化投资与对冲基金丛书”，深受业内好评。在此基础上，2016 年我们再次重磅出击，整合业内顶尖人才，推出“大数据金融丛书”，以引领时代前沿、助力行业发展。

本书特点

我和卓金武认识多年，看到他在业内做得风生水起，这次他的新书《MATLAB 时间序列方法与实践》是一个很有价值的成果。我最初从事金融行业就是从时间序列开始的，那时候我还在上海交通大学当老师，研究的就是利用人工智能技术进行时间序列的分析与预测。时间序列在金融领域的通用说法就是 K 线，所有做技术分析的人士都会对 K 线的走势进行分析，无论是大盘还是个股，或者是期货品种，所有的交易策略，都是在 K 线走势的分析基础之上的。

从学术的角度，研究抽象的时间序列的类型、走势、未来方向，构建通用的模型，毫无疑问，不仅可以用于资本市场，也同样可以用于所有需要时间序列分析的场合，卓金武的这本书在这方面选择了一个非常有价值的方向。

全书可以分为这么几块，第一部分包括第 1～2 章，主要对时间序列做了概念性的描述，读者从中可以了解除 K 线之外，还有更多的与时间有关的数据序列，以及这些序列在实际中的应用。

第二部分包括第 3～10 章，是本书的重点，依次介绍了 AR、MA、ARMA、ARIMA 模型，时间序列平稳性检验、趋势与季节性时间序列建模，ARCH、GARCH 模型和多元时间序列建模。这些模型从不同的角度对时间序列进行解释，并且通过回归、相

关性分析等方法对未来的走势进行一定程度的预测。

第三部分，包括第 11～12 章，介绍了两个时间序列的综合应用实例，一个是关于航空公司的应用，另一个是在股市中的应用。对于大多数读者来说，在股市中的应用可能是他们最关心的，这一部分对大多数的量化投资者来说，是非常有价值的。另外，文中的主要案例，都给出了 MATLAB 实现代码，毫无疑问增加了本书的吸引力。

卓金武的这本书从理论上试图解决时间序列的分类和预测问题，可以说，是从另一个更高的高度来解决目前技术分析体系想要解决的问题，这对于资产管理行业的价值不言而喻，特此推荐。

美好前景

中国经济经过几十年的高速发展，各行各业基本上已经定型，能够让年轻人成长的空间越来越小。未来十年，大数据金融领域是少有的几个有着百倍、甚至千倍成长空间的行业。在传统的以人为主的分析逐步被数据和模型所替代的过程中，从事数据处理、模型分析、交易实现、资产配置的核心人才（我们称之为“宽客”），将有广阔的舞台可以充分展示自己的才华。在这个领域，将不再关心你的背景和资历，无论学历高低，无论有无经验，只要你勤奋、努力、脚踏实地地研究数据、研究模型、研究市场，实现财务自由并非是遥不可及的梦想。对于宽客来说，除了你的才华，其他一切都不重要！

丁鹏　博士

中国量化投资学会　理事长

《量化投资——策略与技术》作者

“大数据金融丛书”主编

2018 年 12 月于上海

前　言

随着对时间序列分析理论与应用两个方面的深入研究，时间序列分析应用的范围日益扩大。目前，它已涉及天文、地理、生物、物理、化学等自然科学领域，图像识别、语音通信、声呐技术、遥感技术、核工程、环境工程、医学工程、海洋工程、冶金工程、机械工程等工程技术领域，国民经济、市场经济、生产管理、人口等社会经济领域，并已取得不少重要应用成果。本书将介绍时间序列分析中的基础知识、常用方法、MATLAB 实现过程和经典的应用案例。

本书内容

全书内容分为 12 章。第 1 章为绪论，介绍时间序列的基本概念和知识体系等基本内容；第 2～10 章为时间序列各具体知识块的详细介绍，并有相关的 MATLAB 实现代码；第 11 章和第 12 章为实践部分。

本书特色

- 知识体系全面，本书涵盖了时间序列学习中所需的基础知识，包括基本的概念、原理、方法和具体实现计算的过程。
- 案例丰富，在书中重要的知识点后，基本都有相关的应用实例，这些案例更直观地描述了时间序列的应用场景和用法，同时加深了对基础概念的理解。
- 详细的 MATLAB 代码，MATLAB 是时间序列分析中功能最强大、应用最广的工具之一，全书的案例都用 MATLAB 实现，代码可以直接借鉴，对实际的时间序列分析会有很大的帮助。

读者对象

- 从事金融、经济、管理、统计的专业人士、教师和学生。
- 从事时间序列研究的科研工作者。
- 希望学习 MATLAB 的工程师或科研工作者。因为本书的代码都是用 MATLAB 编写的，所以对于希望学习 MATLAB 的读者来说也是一本很好的参考书。

致谢

本书的编写、出版得到了中国量化投资学会、电子工业出版社等单位的帮助，在此对这些单位表示感谢。电子工业出版社的李冰老师全程指导本书的编写，在此向她表示感谢！

由于时间仓促，加之作者水平有限，疏漏之处在所难免。在此，诚恳地期待得到广大读者的批评指正。

江渝　李幸　卓金武

2018 年 6 月于上海

目　录

1 绪论

随着对时间序列分析理论与应用两方面研究的不断深入，时序分析应用的范围日益扩大。目前，它已涉及天文、地理、生物、物理、化学等自然科学领域；图像识别、语音通信、声呐技术、遥感技术、核工程、环境工程、医学工程、海洋工程、冶金工程、机械工程等工程技术领域；国民经济、市场经济、生产管理、人口等社会经济领域，并已取得不少重要的应用成果。

1.1 时间序列的发展过程

最早的时间序列分析可以追溯到 7000 年前的古埃及，古埃及人把尼罗河涨落的情况每天记录下来，从而构成了一个时间序列。对这个时间序列长期的观察使他们发现尼罗河的涨落非常有规律，由于掌握了涨落的规律，古埃及的农业迅速发展。这种从观测时间序列得到直观规律的方法即为描述性分析方法。在时间序列分析方法的发展历程中，经济、金融、工程等领域的应用始终起着重要的推动作用，时间序列分析的每一步发展都与应用密不可分。

一般地，人们认为现代时间序列分析起源于英国统计学家 G.u.Yule 在 1927 年提出的 AR（Auto Regressive，自回归）模型。该模型与英国统计学家 G.T.Walker 在 1931 年提出的 MA（Moving Average，滑动平均）模型和 ARMA（Auto Regression Moving Average）模型，共同构成了时间序列分析的基础，至今仍被广泛应用。这三个模型主要应用于单变量、同方差场合的平稳序列。

值得一提的是，Box 和 Jenkins 在 1927 年出版的 *Time Series Analysis: Forecasting and Control* 被认为是时间序列分析发展历程中的里程碑。该书为实际工作者提供了对时间序列进行分析、预测，以及对 ARIMA 模型进行识别、估计和诊断的系统方法。

ARIMA 模型也被称为 Box-Jenkins 模型，主要应用于单变量、同方差场合的线性模型。该模型可以处理非平稳序列，主要思想是先对非平稳序列进行差分，使之变为平稳序列，然后再用 ARMA 模型来拟合差分后的序列。

前面所说的 AR 模型、MA 模型、ARMA 模型和 ARIMA 模型都要求时间序列为单变量、同方差的线性模型。随着时间序列分析及其理论的发展，人们发现这些假设在一些情形下并不成立，例如 Moran（1953）在对加拿大山猫数据的建模过程中发现数据的怪异特征，即大于均值的样本点的残差显著地小于那些小于均值的样本点的残差。因此，人们越来越关心异方差、多变量、非线性的时间序列。

针对异方差情形，Engle（1982）首先提出 ARCH（Auto Regressive Conditional Heteroskedasticity，自回归条件异方差）模型。ARCH 模型的基本思想是假设同一时刻噪声服从均值为零，方差是一个随时间变化的量（即为条件异方差）的正态分布，且这个随时间变化的方差是过去有限项序列值平方的线性组合（即为自回归）。作为一种全新的理论，ARCH 模型在近几十年里得到了极大的发展，已被广泛地应用于验证金融理论中的规律性描述，以及金融市场的预测和决策结果。该模型也被认为是近年来金融计量学发展中最重大的创新。然而，ARCH 模型只适用于异方差函数短期自相关过程，为此 Bollerslev（1986）将 ARCH 模型推广至广义自回归条件异方差（GARCH）模型，GARCH 模型更能反映实际数据的长期记忆性质。ARCH 的另外几种推广形式有 Engle 等人（1987）提出的 ARCH-M 模型和 Nelson（1991）提出的指数广义自回归条件异方差（EGARCH）模型等。

针对多变量的情形，自然的想法是把一维时间序列的分析方法推广至多维。因此，早期多维时间序列的分析方法中，往往要求每个序列都是平稳的。常见的模型有向量 ARMA 模型、向量自回归模型（VAR）等。由于从一元自回归滑动平均模型到多元自回归滑动平均的情形不能直接推广，其中存在很多问题和需要克服的困难，包括模型的识别、估计和解释等，因此这方面的发展相对较慢。直到 Engle 和 Granger（1987）提出了协整（Co-integration）理论及其方法，为多维非平稳序列的建模提供了一种途径。协整理论中，各序列可以都是不平稳的，但它们的线性组合却是平稳序列，该理论可以解释变量之间长期稳定的均衡关系。协整方法已成为了分析线性非平稳序列数量关系的最重要工具之一。对于序列之间存在非线性调整机制的情形，Balke 和

Fomby（1997）提出了阈值协整（Threshold Cointegration）方法。例如，在股票交易过程中，由于交易费用、交易政策等因素的变化会导致股价的非对称调整；国家的货币政策由于制度方面的原因也会使通货膨胀率产生非对称调整。

针对非线性情形，Tong 和 Lim（1980）提出了 TAR（Threshold Autoregressive Regressice，门限自回归）模型。TAR 模型假定在状态空间的不同区域，模型有不同的线性形式，状态空间的划分通常由一个门限变量来确定，该模型属于参数模型。近二十年来，人们更多地关注时间序列的非参数模型，如非参数自回归（NAR）模型、非参数自回归异方差（NARCH）模型等。

时间序列分析方法的另一个突破是在谱分析方面。给定一个时间序列样本，通过傅里叶变化可以把时域上的数据变换到频域，这就是经典谱分析方法，例如周期图谱法等。Burg（1967）在他从事的地震信号分析与处理的工作中提出最大熵谱，把信息熵的概念融入信号处理中，有时又称为时序谱分析方法，是现代谱分析的开始。Capon（1969）提出了最小方差谱估计方法。这两个方法共同奠定了现代谱估计的基础。此后 Shore 和 Johnson（1980）又提出了最小交叉熵法。理论证明，最大熵谱分析法只是最小交叉熵法的一个特例。当存在先验信息时，最小交叉熵法可获得比最大熵法好得多的分辨率。但最小交叉熵法的缺点是运算太复杂。一般地，经典谱分析对于长数据序列有良好的谱估计性能，但对于短数据序列，经典谱分析存在分辨率不高等致命弱点，现代谱估计法则具有优良性能。

1.2 时间序列的基本概念

1.2.1 时间序列的定义

所谓时间序列就是一组按照一定的时间间隔排列的一组数据，其时间间隔可以是任意的时间单位，如小时、日、周、月等。这一组数据可以表示各种各样的含义，如经济领域中每年的产值、国民收入、商品在市场上的销量、股票数据的变化情况等；社会领域中某一地区的人口数、医院患者人数、铁路客流量等；自然领域的太阳黑子数、月降水量、河流流量等，这些数据都形成了一个时间序列。人们希望通过对这些时间序列的分析，从中发现和揭示现象的发展和变化规律，或从动态的角度描述某一

现象和其他现象之间的内在数量关系及其变化规律，从而尽可能多地从中提取出所需要的准确信息，并将这些知识和信息用于预测，以掌握和控制未来行为。人们研究时间序列，通常也是希望根据历史数据预测未来的数据。对于时间序列的预测，由于很难确定它与其他因变量的关系，或收集因变量的数据非常困难，这时就不能采用回归分析方法进行预测，而是需要使用时间序列分析方法来进行预测。

采用时间序列分析进行预测时需要用到一系列的模型，这种模型统称为时间序列模型。在使用这种时间序列模型时，总是假定某一种数据变化模式或某一种组合模式会重复发生。因此首先需要识别出这种模式，然后采用外推的方式进行预测。采用时间序列模型进行分析时，显然其关键在于辨识数据的变化模式（样式）；同时，决策者所采取的行动对这个时间序列的影响很小，因此这种方法主要用来对一些环境因素，或不受决策者控制的因素进行预测，如宏观经济情况、就业水平、某些产品的需求量等数据。

这种方法的主要优点是数据很容易得到，而且容易被决策者理解，计算相对简单。当然对于高级时间序列分析法，其计算也是非常复杂的。此外，时间序列分析法常常用于中短期预测，因为在相对短的时间内，数据变化的模式不会特别显著。

时间序列分析的主要用途有：①系统描述。根据对系统进行观测，得到时间序列数据，用曲线拟合方法对系统进行客观的描述。②系统分析。当观测值取自两个以上变量时，可用一个时间序列中的变化去说明另一个时间序列中的变化，从而深入了解给定时间序列产生的机理。③预测未来。一般用 ARMA 模型拟合时间序列，预测该时间序列的未来值。④决策和控制。根据时间序列模型可调整输入变量，使系统发展过程保持在目标值上，即预测到要偏离目标时，便可进行必要的控制。

1.2.2 时间序列的组成因素

时间序列的变化受许多因素的影响，有些起着长期的、决定性的作用，使其呈现出某种趋势和一定的规律性；有些则起着短期的、非决定性的作用，使其呈现出某种不规则性。在分析时间序列的变动规律时，事实上不可能将每个影响因素都一一划分开来，分别去作精确分析，但可以将众多影响因素，按照对现象变化影响的类型，划分成若干时间序列的构成因素，然后对这几类构成要素分别进行分析，以揭示时间序列的变动规律性。影响时间序列的构成因素可归纳为以下 4 种。

（1）趋势性（Trend），指现象随时间推移朝着一定方向呈现出持续渐进的上升、下降，平稳的变化或移动。这一变化通常是许多长期因素的结果。

（2）周期性（Cyclic），指时间序列表现为循环于趋势线上方和下方的点序列，并持续一段时间以上的有规则变动。这种因素具有周期性的变动，比如高速通货膨胀时期后面紧接的温和通货膨胀时期，将会使许多时间序列表现为交替地出现于一条总体递增趋势线的上下方。

（3）季节性变化（Seasonal Variation），指现象受季节性影响，按一固定周期呈现出的周期波动变化。尽管通常将一个时间序列中的季节变化认为是以 1 年为期的，但是季节因素还可以被用于表示时间长度小于 1 年的有规则重复形态。比如，每日交通量数据表现出为期 1 天的“季节性”变化，即高峰期到达高峰水平，而一天的其他时期车流量较小，从午夜到次日清晨最小。

（4）不规则变化（Irregular Movement），指现象受偶然因素的影响而呈现出的不规则波动。这种因素包括实际时间序列值与考虑了趋势性、周期性、季节性变动的估计值之间的偏差，它用于解释时间序列的随机变动。不规则因素是由短期的未被预测到的，以及不被重复发现的那些影响时间序列的因素引起的。

时间序列一般是上述几种变化形式的叠加或组合，如图 1-1 所示。

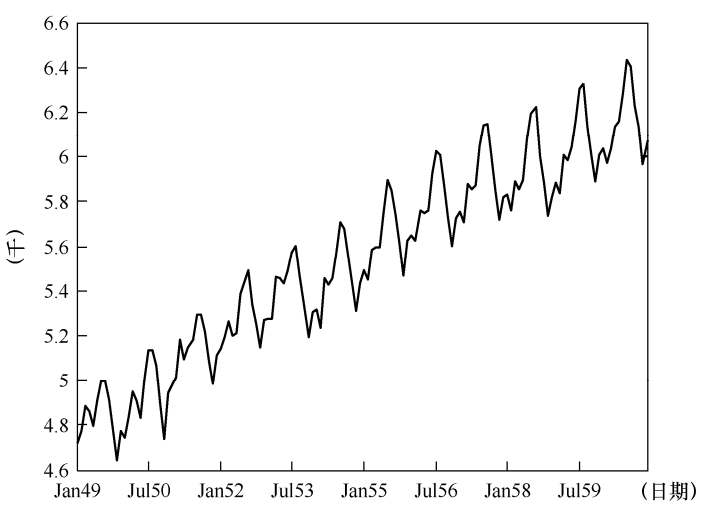


图 1-1　一种时间序列的叠加形式

1.2.3 时间序列的分类

根据不同的标准，时间序列有不同的分类方法，常用的标准及分类方法如下。

（1）按所研究的对象的多少来分，有一元时间序列和多元时间序列，如某种商品的销售量数列，即为一元时间序列；如果所研究对象不仅仅是一个数列，而是多个变量，如按年、月顺序排序的气温、气压、雨量数据等，每个时刻对应着多个变量，则这种序列为多元时间序列。

（2）按时间的连续性，可将时间序列分为离散时间序列和连续时间序列两种。如果某一序列中的每一个序列值所对应的时间参数为间断点，则该序列就是一个离散时间序列；如果某一序列中的每个序列值所对应的时间参数为连续函数，则该序列就是一个连续时间序列。

（3）按序列的统计特性，分为平稳时间序列和非平稳时间序列两类。所谓时间序列的平稳性，是指时间序列的统计规律不会随着时间的推移而发生变化。平稳序列的时序图直观上应该显示出该序列始终在一个常数值附近随机波动，而且波动的范围有界、无明显趋势及无周期特征。相对的，时间序列的非平稳性，是指时间序列的统计规律随着时间的推移而发生变化。

（4）按序列的分布规律来分，有高斯型（Guassian）和非高斯型（Non-Guassian）时间序列两类。

1.2.4 时间序列分析方法

时间序列分析是一种被广泛应用的数据分析方法，它研究的是代表某一现象的一串随时间变化而又相互关联的数字系列（动态数据），从而描述和探索该现象随时间发展、变化的规律性。时间序列分析利用的手段可以通过直观简便的数据图法、指标法、模型法等来分析。而模型法相对来说更具体也更深入，能更本质地了解数据的内在结构和复杂特征，以达到控制与预测的目的。总的来说，时间序列分析方法包括如下两类。

（1）确定性时序分析：指暂时过滤掉随机性因素（如季节因素、趋势变动）进行确定性分析的方法，其基本思想是用一个确定的时间函数 $y = f(t)$ 来拟合时间序列，不同的变化采取不同的函数形式来描述，不同变化的叠加采用不同的函数叠加来描

述。具体可分为趋势预测法（最小二乘法）、平滑预测法、分解分析法等。

（2）随机性时序分析：其基本思想是通过分析不同时刻变量之间的相关关系，揭示其相关结构，利用这种相关结构建立自回归、滑动平均、自回归滑动平均混合模型来对时间序列进行预测。

无论采用哪种方法，时间序列的一般的分析流程基本固定，如图 1-2 所示。

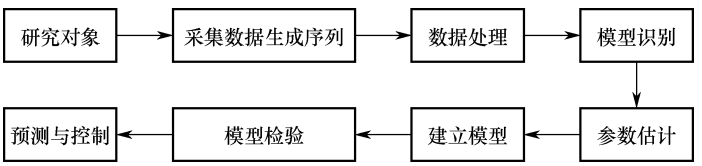


图 1-2　时间序列分析流程

1.3　平稳时间序列分析方法

时间序列的变动是长期趋势变动、季节变动、循环变动、不规则变动的耦合或叠加。在确定性时间序列分析中通过移动平均、指数平滑、最小二乘法等方法来体现出社会经济现象的长期趋势及带季节因子的长期趋势，预测未来的发展趋势。

1.3.1　移动平均法

1．一次移动平均法

一次移动平均法指收集一组观察值，计算这组观察值的均值，并利用这一均值作为下一期的预测值的预测方法。其模型为：

$$M_t^{(1)} = \frac{X_t + X_{t-1} + \cdots + X_{t-N+1}}{N}$$

$$\hat{X}_{t+1} = M_t^{(1)}$$

式中，X_t 为 t 期的实际值；N 为所选数据个数；$\hat{X}_{t+1}$ 为下一期(t+1)的预测值。

2．二次移动平均法

二次移动平均法的线性模型为：

$$\hat{X}_{t+1} = a_t + b_t T$$

$$M_t^{(1)} = \frac{X_t + X_{t-1} + \cdots + X_{t-N+1}}{N}$$

$$M_t^{(2)} = \frac{M_t^{(1)} + M_{t-1}^{(1)} + \cdots + M_{t-N+1}^{(1)}}{N}$$

$$a_t = 2M_t^{(1)} - M_t^{(2)}$$

$$b_t = \frac{2(M_t^{(1)} - M_t^{(2)})}{N-1}$$

式中，X_t 为 t 期的实际值；$\hat{X}_{t+1}$ 为(t+1)期的预测值；t 为当前的时期数；T 为由 t 至预测期的时期数。

采用移动平均法进行预测，用来求平均数的时期数 N 的选择非常重要，这也是移动平均的难点。因为 N 取值的大小对所计算的平均数的影响较大。当 N=1 时，移动平均预测值为原数据的序列值；当 N=全部数据的个数时，移动平均值等于且为全部数据的算术平均值。显然，N 值越小，表明对近期观测值预测的作用越重视，预测值对数据变化的反应速度也越快，但预测的修匀程度较低，估计值的精度也可能降低；反之，N 值越大，预测值的修匀程度越高，但对数据变化的反应速度较慢。

不存在一个确定时期 N 值的规则。一般 N 在 2～200 之间，视序列长度和预测目标情况而定。一般对水平型数据，N 值的选取较为随意。一般情况下，如果考虑到历史上序列中含有大量随机成分，或者序列的基本发展趋势变化不大，则 N 应取大一点。对于具有趋势性或阶跃性特点的数据，为提高预测值对数据变化的反应速度，减少预测误差，N 值应取得较小一些，以使移动平均值更能反映目前的发展变化趋势。一般 N 的取值为 2～15，具体取值要根据实际情况来定。

1.3.2 指数平滑法

1．一次指数平滑法

一次指数平滑法的基本模型为：

$$S_t^{(1)} = \alpha X_t + (1-\alpha) S_{t-1}^{(1)}$$

或

$$S_t^{(1)} = \alpha X_t + \alpha(1-\alpha)X_{t-1} + \cdots + \alpha(1-\alpha)^{t-1}X_1 + (1-\alpha)^t S_0^{(1)}$$

下一期的预测值为：

$$\hat{X}_{t+1} = S_t^{(1)}$$

式中，$X_0, X_1, \cdots, X_n$ 为时间序列观测值；$S_0^{(1)}$，$S_1^{(1)}, \cdots, S_n^{(1)}$ 为观测值的指数平滑值；α 为平滑系数（$0<\alpha<1$）。

一次指数平滑法比较简单，但必须设法找到最佳的 α 值，以使均方差最小，这需要通过反复试验才能确定。

2．二次指数平滑法

二次指数平滑法的线性模型为：

$$\hat{X}_{t+T} = a_t + b_t T$$

$$a_t = 2S_t^{(1)} - S_t^{(2)}$$

$$b_t = \frac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)})$$

$$S_t^{(1)} = \alpha X_t + (1-\alpha)S_{t-1}^{(1)}$$

$$S_t^{(2)} = \alpha S_t^{(1)} + (1-\alpha)S_{t-1}^{(2)}$$

式中，$S_t^{(1)}$、$S_t^{(2)}$ 分别是一次指数平滑值和二次指数平滑值；X_t 为 t 期的实际值；$\hat{X}_{t+T}$ 为 $t+T$ 期的预测值；α 为平滑系数（$0<\alpha<1$）。

1.4 季节指数预测法

季节指数法是指变量在一年内以（季）月的循环为周期特征，通过计算季节指数达到预测目的的一种方法。其操作过程为：首先分析判断时间序列数据是否呈现季节性波动，一般将 3～5 年的资料按（季）月展开，然后绘制历史曲线图，观察其在一年内有无周期性波动来作判断。在下面的讨论中，设时间序列数据为 X_1，X_2,…，X_{4n}，

n 为年数，每年取 4 个季度。

1.4.1 季节性水平模型

如果时间序列没有明显的趋势变动，而主要受季节变化和不规则变动影响时，可用季节性水平模型进行预测。预测模型的方法如下。

1．计算历年同季的平均数

$$\begin{cases} r_1 = \dfrac{1}{n}(X_1 + X_5 + \cdots + X_{4n-3}) \\ r_2 = \dfrac{1}{n}(X_2 + X_6 + \cdots + X_{4n-2}) \\ r_3 = \dfrac{1}{n}(X_3 + X_7 + \cdots + X_{4n-1}) \\ r_4 = \dfrac{1}{n}(X_4 + X_8 + \cdots + X_{4n}) \end{cases}$$

2．计算全季总平均数

$$y = \frac{1}{4n}\sum_{i=1}^{4n} X_i$$

3．计算各季的季节指数

历年同季的平均数与全时期的季平均数之比，即：

$$\alpha_i = \frac{r_i}{y}(i = 1,2,3,4)$$

若各季的季节指数之和不为 4，季节指数需要调整为：

$$F_i = \frac{4}{\sum \alpha_i}\alpha_i(i = 1,2,3,4)$$

4．利用季节指数法进行预测

$$\hat{X}_t = X_i \frac{\alpha_t}{\alpha_i}$$

式中，$\hat{X}_t$ 为第 t 季的预测值；α_t 为第 t 季的季节指数；X_i 为第 i 季的实际值；α_i 为第 i 季的季节指数。

1.4.2 季节性趋势模型

当时间序列既有季节性变动又有趋势性变动时，先建立季节性趋势预测模型，在此基础上求得季节指数，再建立预测模型，其过程如下。

（1）计算历年同季平均数 r。

（2）建立趋势预测模型求趋势值 $\hat{X}_y$，直接用原始数据时间序列建立线性回归模型即可。

（3）计算出趋势值后，再计算出历年同季的平均值 R。

（4）计算趋势季节指数 k，用同季平均数与趋势值同季平均数之比来计算。

（5）对趋势季节指数进行修正。

（6）求预测值，将预测值的趋势值乘以该期的趋势季节指数，即预测模型为：

$$\hat{X}_t = k\hat{X}_y$$

1.5 时间序列主要模型介绍

1.5.1 ARMA 模型

ARMA 模型的全称是自回归移动平均（Auto Regression Moving Average）模型，它是目前最常用的来拟合平稳时间序列的模型。ARMA 模型又可细分为 AR 模型、MA 模型和 ARMA 模型三大类。

1．AR（p）（p 阶自回归模型）

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + u_t$$

其中，u_t 是白噪声序列；δ 是常数（表示序列数据没有 0 均值化）。

2．MA（q）（q 阶移动平均模型）

$$X_t = \mu + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}$$

其中$\{u_t\}$是白噪声过程；MA（q）是由 u_t 本身和 q 个 u_t 的滞后项加权平均构造出来的，因此它是平稳的。

3．ARMA（p, q）（自回归移动平均过程）

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + \delta + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}$$

其中的参数含义同 AR、MA 模型，ARMA 模型相当于 AR 模型和 MA 模型的叠加。

1.5.2 ARIMA 模型

ARIMA 模型全称为差分自回归移动平均模型（Autoregressive Integrated Moving Average Model），是由博克思（Box）和詹金斯（Jenkins）于 20 世纪 70 年代初提出的著名时间序列预测方法，所以又称为 Box-Jenkins 模型、博克思—詹金斯法。ARIMA 模型是 ARMA 模型的拓展，可以表示为 ARIMA（p，d，q），其中 AR 是自回归，p 为自回归项； MA 为移动平均，q 为移动平均项数；d 为时间序列成为平稳序列时所做的差分次数。所谓 ARIMA 模型，是指将非平稳时间序列转化为平稳时间序列，然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归，所建立的模型。ARIMA 模型根据原序列是否平稳，以及回归中所含部分的不同，包括移动平均过程（MA）、自回归过程（AR）、自回归移动平均过程（ARMA），以及 ARIMA 过程。

ARIMA 模型的基本思想是：将预测对象随时间推移而形成的数据序列视为一个随机序列，用一定的数学模型来大概描述这个序列。这个模型一旦被识别后就可以利用时间序列的过去值及现在值来预测未来值。

由于 ARIMA 模型是 ARMA 模型的拓展，ARIMA 包含 ARMA 模型的三种形式，即 AR、MA、ARMA 模型，另外它还有一种经差分的 ARMA 模型形式，即：

$$\Delta x_t = x_t - x_{t-1} = x_t - Lx_t = (1-L)x_t$$
$$\Delta^2 x_t = \Delta x_t - \Delta x_{t-1} = (1-L)x_t - (1-L)x_{t-1} = (1-L)^2 x_t$$
$$\Delta^d x_t = (1-L)^d x_t$$

对于 d 阶单整序列 $I(d)$，令：

$w_t = \Delta^d X_t = (1-L)^d X_t$，则 w_t 是平稳序列。于是可对 w_t 建立 ARMA（p，q）模型，所得到的模型称为 X_t～ARIMA（p，d，q）模型，故 ARIMA（p，d，q）模型可以表示为：

$$w_t = \varphi_1 w_{t-1} + \varphi_2 w_{t-2} + \cdots + \varphi_p w_{t-p} + \delta + u_t + \theta_1 u_{t-1} + \theta_2 u_{t-2} + \cdots + \theta_q u_{t-q}$$

1.5.3 ARCH 模型

ARCH 模型（Autoregressive Conditional Heteroskedasticity Model）全称为“自回归条件异方差模型”，由罗伯特·恩格尔在 1982 年发表在《计量经济学》杂志（Econometrica）的一篇论文中首次提出。ARCH 模型解决了时间序列的波动性（Volatility）问题，这个模型是获得 2003 年诺贝尔经济学奖的计量经济学成果之一。目前该模型已被认为集中地反映了方差的变化特点，从而广泛地被应用于经济领域的时间序列分析。

ARCH 模型的定义：若一个平稳随机变量 X_t 可以表示为 AR（p）形式，其随机误差项的方差可用误差项平方的 q 阶分布滞后模型描述，即

$$X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_{p1} X_{t-p} + u_t \qquad \text{（a）}$$

$$\sigma_t^2 = E(u_t^2) = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_q u_{t-q}^2 \qquad \text{（b）}$$

则称 u_t 服从 q 阶的 ARCH 过程，记作 u_t~ARCH（q）。其中（a）式称作均值方程，（b）式称作 ARCH 方程。

ARCH 模型经常以回归的方式来描述，这也是常见的 ARCH 模型的另一种描述方式：

$$\begin{cases} X_t = c + \rho_1 X_{t-1} + \rho_2 X_{t-2} + \cdots + \rho_m X_{t-m} + \varepsilon_t \\ \varepsilon_t = \sqrt{h_t} v_t \\ h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 \end{cases} \qquad \text{（c）}$$

其中 v_t 服从独立同分布；式（c）和上面第一种描述是等价的，但（c）式的操作性更强。

1.5.4 GARCH 模型

GARCH 模型称为广义 ARCH 模型，是 ARCH 模型的拓展，由 Bollerslev（1986）提出。

GARCH（p，q）的模型可表示为：

$$\begin{cases} X_t = c + \rho_1 X_{t-1} + \rho_2 X_{t-2} + \cdots + \rho_m X_{t-m} + \varepsilon_t \\ \varepsilon_t = \sqrt{h_t} v_t \\ h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i} \end{cases}$$

GARCH 模型实际上就是在 ARCH 模型的基础上，考虑了异方差函数的 p 阶自相关性而形成的，它可以有效拟合具有长期记忆的异方差函数。显然 ARCH 模型是 GARCH 模型的一个特例，ARCH（q）模型实际上就是 p=0 时的 GARCH（p，q）模型。

1.6 时间序列分析工具

表 1-1 给出了 5 种常用的时间序列分析工具在功能、特点、适应情况方面的比较，从中可以看出，5 款工具都具有自己独特的特点，都有一定的适应条件。

表 1-1 常用时间序列分析工具的比较

名　称	功　能	特　点	适应情况
MATLAB	时间序列分析功能主要依靠其计量经济学工具箱，具有强大、灵活的科学计算能力	1）擅长矩阵计算和仿真模拟 2）丰富的数学函数，适合算法开发或自主的程序开发 3）强大的绘图功能	适合于模型研究、开发产品、频发的数据交互、涉及其他科学计算等情况
Eviews	专业的计量经济学分析工具，包含主要的时间序列分析模型和相关的检验方法	1）界面操作 2）不需要写代码就能实现基本的时间序列分析 3）灵活性不够，不便于组合分析	时间序列模型明确，只需要得到情况

续表

名　称	功　能	特　点	适应情况
SAS	功能极强大的统计分析软件，具有时间序列分析功能	1）较强的大数据处理能力 2）支持二次开发	有一些行业标准，适合工业使用
SPSS	侧重统计分析，具有时间序列分析功能	SPSS 使用方便，但不适合自己开发代码，就是说扩展上受限，如果要求不高，已是足够了	界面友好，使用简单，但是功能很强大，也可以编程，能解决绝大部分统计学问题，适合初学者
R	类似 MATLAB，具有丰富的数学和统计分析函数	R 是开源的，支持二次开发	适合于算法学习、产品研发，小项目的开发

纵观这 5 种工具的特点，本书将选择 MATLAB 作为时间序列分析主要的实现工具。主要有三个方面的原因：一是因为时间序列的主要内容是各种各样的模型，而 MATLAB 特别适合高效自主的模型开发，因为 MATLAB 具有丰富的数学函数库，可以使用这些函数库快速实现模型，从而加强对模型的学习和理解；二是因为 MATLAB 具有丰富的科学计算功能，包括微积分、优化计算、符号计算等，以及丰富的金融工具和经济工具箱；三是 MATLAB 本身就是程序开发工具，具有 GUI 界面开发功能，所以使用 MALTAB 可以很快将学习的模型开发成程序和工具，部署到实际的应用环境中。

工具都是相通的，只要掌握一种工具后，再去应用其他工具，很快就会上手，本书只是从学习的角度，以为 MATLAB 更合适些。

1.7　应用实例：基于时间序列的股票预测

有些股票的价格波动具有很好的周期性，这时就可以考虑用时间序列方法进行股票价格预测。下面将以具体的实例来说明如何利用上述介绍的时间序列方法进行股票价格走势的预测。

（1）读取股票数据

```
    clc, clear all, close all
Y=xlsread('sdata','Sheet1','E1:E227');
N = length(Y);
```

（2）原始数据可视化

```
    figure(1)
plot(Y); xlim([1,N])
set(gca,'XTick',[1:18:N])
title('原始股票价格')
ylabel('元')
```

该节程序执行后，会得到如图 1-3 所示的原始的股票价格走势图，从该图中可以看出，股票的价格变动有些规律，即周期性上升，为此可以考虑用时间序列来建立股票价格走势的模型。

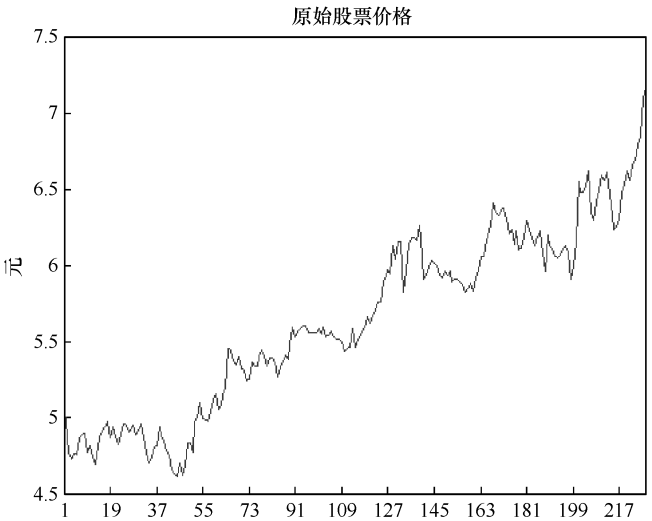


图 1-3　原始的股票价格走势图

（3）建立 ARIMA 模型

由于 ARIMA 具有较强的适应性，可以尝试用该模型建立该股票的时间序列模型，具体代码如下：

```
    model = arima('Constant',0,'D',1,'Seasonality',12,...
            'MALags',1,'SMALags',12)
Y0 = Y(1:13);
[fit,VarCov] = estimate(model,Y(14:end),'Y0',Y0);
```

代码执行后，得到如下 ARIMA 模型参数：

```
model =
```

```
ARIMA(0,1,1) Model Seasonally Integrated with Seasonal MA(12):
------------------------------------------------------------
Distribution: Name = 'Gaussian'
          P: 13
          D: 1
          Q: 13
   Constant: 0
         AR: {}
        SAR: {}
         MA: {NaN} at Lags [1]
        SMA: {NaN} at Lags [12]
 Seasonality: 12
   Variance: NaN

ARIMA(0,1,1) Model Seasonally Integrated with Seasonal MA(12):
------------------------------------------------------------
Conditional Probability Distribution: Gaussian
                          Standard          t
Parameter      Value         Error      Statistic
-----------   -----------   ------------   -----------
Constant            0         Fixed         Fixed
   MA{1}     0.0654479    0.0706347      0.926568
  SMA{12}     -0.78655    0.0370049      -21.2553
Variance    0.00972519   0.000703112       13.8316
```

（4）评估预测效果

```
   Y1 = Y(1:100);
Y2 = Y(101:end);

Yf1 = forecast(fit,100,'Y0',Y1);

figure(2)
plot(1:N,Y,'b','LineWidth',2)

hold on
plot(101:200,Yf1,'k--','LineWidth',1.5)
xlim([0,200])
```

```
title('Prediction Error')
legend('Observed','Forecast','Location','NorthWest')
hold off
```

该节程序运行后，产生如图 1-4 所示的股票实际走势与预测走势的比较图。从该图中可以看出，两者总的趋势一致，但波动周期、波动幅度差异较大。这说明时间序列能在一定程度上反映股价的走势情况，但同时也说明现实中股价的变化情况具有较强的无序、随机的特征。这也是比较客观的，因为时间序列模型是经过抽象后形成的比较完美的模型，而现实世界的股价则是完全自由的，用完美、固定的模型只能刻画现实数据的部分特征。

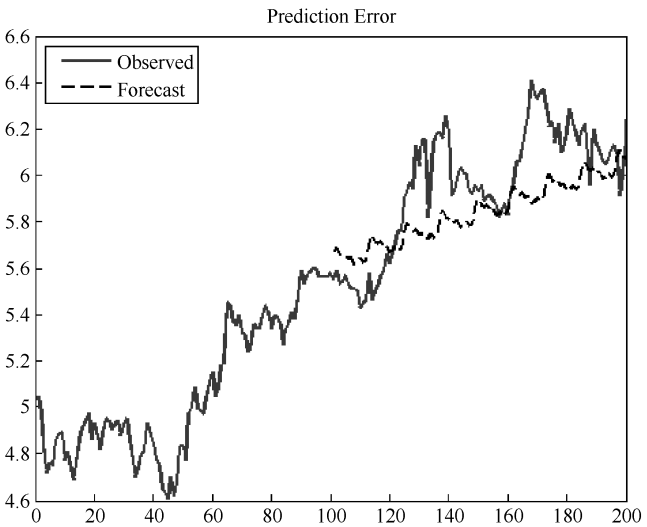


图 1-4　股票实际走势与预测走势的比较图

（5）预测未来股票趋势

```
[Yf,YMSE] = forecast(fit,60,'Y0',Y);
upper = Yf + 1.96*sqrt(YMSE);
lower = Yf - 1.96*sqrt(YMSE);

figure(3)
plot(Y,'b')
hold on
h1 = plot(N+1:N+60,Yf,'r','LineWidth',2);
h2 = plot(N+1:N+60,upper,'k--','LineWidth',1.5);
```

```
plot(N+1:N+60,lower,'k--','LineWidth',1.5)
xlim([0,N+60])
title('95%置信区间')
legend([h1,h2],'Forecast','95% Interval','Location','NorthWest')
hold off
```

本节程序得到的是用已经训练的模型对未来股价预测后的结果，如图 1-5 所示，同时还得到股价 95%的置信波动区间，说明股价的可能波动范围。从该图中可以看出，预测时间越长，波动范围越大，这也说明预测时间越长，结果越不准，所以在用时间序列预测时，尽量不要将预测时间设置得太长，原则上预测时间不宜超过时间序列数据对应总时长的 10%，也就是向后推延的时间不超过历史时间的 10%。

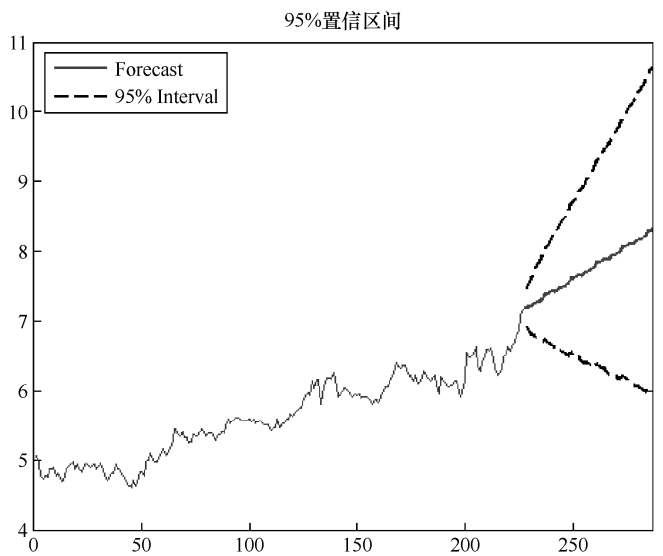


图 1-5　对未来股价预测后的结果

从该案例读者也可以体会到，股价数据随机性较强、噪声偏多，时间序列方法可在一定程度上反映股价的走势，对投资具有一定的指导意义。同时需要注意的是，影响股价的因素很多，各种各样非市场的因素往往左右着股价的整体走势，这在一个成熟市场是不应该出现的，从而充分地说明了我国股市还存在一些弊端。对广大投资者而言，要努力提高自身素质，减少对股票的盲目侥幸的认识，培养起应有的投资意识；对股市的研究人员而言，应该积极吸收西方发达国家成熟股市的先进经验和理论，运用于我国股票市场，以起到理论带动实践发展的作用。

1.8 小结

本章对时间序列的发展历程、基本概念、分析方法、主要模型、分析工具等内容进行了介绍，主要是让读者了解时间序列这门课程的全貌。介绍的例子也是让读者知道在实践中如何去使用时间序列方法得到期望的结果。还需要注意的是，一个典型的时间序列分析的过程，首先需要判断时序的类型，对于平稳时间序列则可以用移动平均、指数平滑等方法；如果带有明显的季节特征，则可以用季节指数预测法。对于非平稳时间序列，则需要借助经典的模型进行分析，典型的就是 ARIMA 和 GARCH 两类模型。本章介绍的是时间序列的全貌，后面的章节将展开具体的内容，详细介绍各个具体的模型。

参考文献

[1] 高铁梅. 计量经济分析方法与建模. 北京：清华大学出版社，2006.

[2] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[3] R.F. Engle (1982)，Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation，Econometrica，50，987-1008.

[4] T. Bollerslev (1986)，A Generalized Autoregressive Conditional Heteroskedasticity，Journal of Econometrics 31，307-27.

[5] R.F. Engle, D.M. Lilien, and R.P. Robins (1987), Estimating time varying risk premia in the term structure: the ARCH-M model，Econometrica 55，391-407.

2 时间序列基本概念

时间序列（Time Series）是一组按照时间顺序排列的随机变量，理论研究中经常将其理解为一个随机过程。随机过程（Stochastic Process）是一组有序的随机变量，可以记为$\{\boldsymbol{X}_t, t\in T\}$。随机过程一般是定义在连续集合上的，通常称定义在离散集合上的随机过程为时间序列。一组离散的时间集合 $\boldsymbol{T}$ 可以表示为$\boldsymbol{T}=\{\cdots,-2,-1,0,1,2,\cdots\}$，此时 X_t 是离散时间 t 的随机函数，时间序列通常表示为$\{X_t, t=\cdots,-2,-1,0,1,2,\cdots\}$。

时间序列在特定时间段上的观测样本可以视为随机过程的一次实现，通常称为样本序列，记为$\{X_0, X_1, X_2, \cdots, X_T\}$。理论上说，时间序列可以有无限个观测时间点，然而从实际可获得的样本数据来看，样本序列都是有限的。时间序列经验研究的一个显著特点是，只能在唯一可观测到的样本序列的基础上来推测时间序列的总体特性。

为简便起见，本书中有些时候将随机过程简称为过程，有些时候将时间序列简称为序列，都表示为$\{X_t\}$。

2.1 时间序列的统计概念

我们简要地回顾一下统计分布的一些基本性质和随机变量的矩，设R^k表示 k–维欧几里得空间，$x\in\mathrm{R}^k$表示 x 是R^k的中的点，考虑两个随机向量$X=(X_1,\cdots,X_k)'$和$Y=(Y_1,\cdots,Y_q)'$。令$\mathbf{P}=(X\in\mathrm{A}, Y\in\mathrm{B})$表示$X$在子空间$\mathrm{A}\subset\mathrm{R}^k$中，且$Y$在子空间$\mathrm{B}\subset\mathrm{R}^q$中的概率，由此可以得到这两个随机向量的一些统计概念。

2.1.1 联合分布函数

参数为θ的X与Y的联合分布表示为：

$$F_{X,Y}(x, y, \theta)=P(X\leqslant x, Y\leqslant y;\theta)$$

其中不等号“≤”是分量对分量的运算；X 和 Y 的规律由 $F_{X,Y}(x,y,\theta)$ 刻画。如果 X 和 Y 的联合概率密度函数 $f_{x,y}(x,y,\theta)$ 存在，则：

$$F_{X,Y}(x,y,\theta)=\int_{-\infty}^{x}\int_{-\infty}^{y}f_{x,y}(w,z;\theta)dzdw$$

这时，X 和 Y 是连续型随机向量。

2.1.2 边际分布

X 的边际分布是：

$$F_X(x;\theta)=F_{X,Y}(x,\infty,\theta)$$

这样，X 的边际分布可通过对 Y 求积分得到，同理，Y 的边际分布也可类似得到。

如果 k=1，X 是一个一元随机变量，其分布函数为：

$$F_X(x)=P(X\leqslant x;\theta)$$

称为 X 的累积分布函数（Cumulative Distribution function,CDF）。一个随机变量的 CDF 是非降的（即对 $x_1\leqslant x_2$ 有 $F_X(x_1)\leqslant F_X(x_2)$，且有 $F_X(-\infty)=0$）。

2.1.3 概率分布

对于时间序列 $\{X_t,\ t\in T\}$,我们这样来定义它的概率分布：任取正整数 m，任取 $t_1,t_2,\cdots,\ t_m\in T$，则 m 维随机向量 $(X_{t_1},X_{t_2},\cdots X_{t_m})$ 的联合概率分布记为：

$$F_{t_1,t_2,\cdots,t_m}(X_{t_1},X_{t_2},\cdots X_{t_m})$$

由这些有限维分布函数构成的全体：

$$\left\{F_{t_1,t_2,\cdots,t_m}(X_{t_1},X_{t_2},\cdots X_{t_m}),\forall \mathrm{m}\in \text{正整数},\forall t_1,t_2,\cdots,t_m\in T\right\}$$

就称为序列$\{X_t\}$的概率分布族。

2.1.4 特征统计量

一种更为简单，更实用的描述时间序列统计特征的方法是研究该序列的低阶矩，

特别是均值、方差、自协方差和自相关系数，它们也称为特征统计量。

尽管这些特征统计量并不能描述出随机序列的主要概率特征，但是由于它们的概率意义明显，易于计算，而且往往能代表随机序列的主要概率特征，所以我们对时间序列进行分析，主要就是通过分析这些特征量的统计特征，推断出随机序列的性质。

1．均值

对时间序列$\{X_t, t\in T\}$而言，任意时刻的序列值X_t都是一个随机变量，都有它自己的概率分布，记X_t的分布函数为$F_t(X)$。只要满足条件：

$$\int_{-\infty}^{\infty} X dF_t(X) < \infty$$

就一定存在某个常数μ_t，使得随机变量X_t总是围绕在常数值μ_t附近做随机波动。我们称μ_t为序列$\{X_t\}$在t时刻的均值函数：

$$\mu_t = EX_t = \int_{-\infty}^{\infty} X dF_t(X)$$

当t取遍所有的观察时刻时，就得到一个均值函数序列$\{\mu_t, t\in T\}$。它反映的是时间序列$\{X_t, t\in T\}$每时每刻的平均水平。

2．方差

当$\int_{-\infty}^{\infty} X^2 dF_t(X) < \infty$时，可以将时间序列的方差函数定义为用以描述序列值围绕其均值做随机波动时的平均波动程度：

$$\sigma_t^2 = DX_t = E(X_t - \mu_t)^2 = \int_{-\infty}^{\infty} (X_t - \mu_t)^2 dF_t(X)$$

同样，当t取遍所有的观察时刻时，我们得到的一个方差函数序列$\{\sigma_t^2, t\in T\}$。

3．自协方差函数

类似于协方差函数和相关系数的定义，在时间序列分析中我们定义自协方差函数（Autocovariance Function）和自相关系数（Autocorrelation Function）的概念。

对于时间序列$\{X_t, t\in T\}$，任取$t,s\in T$，定义$\gamma(t,s)$为序列$\{X_t\}$的自协方差函数：

$$\gamma(t,s) = \mathrm{Cov}(X_t, X_s) = E(X_t - \mu_t)(X_s - \mu_s)$$

考虑时间序列 X_t 与它的过去值 X_{t-l} 的线性相关性时，可以把相关系数的概念推广到自相关系数。X_t 与 X_{t-l} 的相关系数称为 X_t 的间隔为 l 的自相关系数，通常记为 ρ_l，在弱平稳性的假定下它只是 l 的函数。定义 $\rho(t,s)$ 为时间序列$\{X_t\}$的自相关系数，简记为 ACF。

$$\rho(t,t-l)=\frac{\gamma(t,t-l)}{\sqrt{DX_t \bullet DX_{t-l}}}$$

之所以称它们为自协方差函数和自相关系数是因为通常的协方差函数和相关系数度量的是两个不同的事件彼此之间的相互影响程度；而自协方差函数和自相关系数度量的是同一事件在两个不同时期之间的相关程度，形象地讲就是度量自己过去的行为对自己现在的影响。

2.2 时间序列的平稳性

在进行时间序列分析时，针对时间序列的平稳（Stationary）和非平稳特性需要采取不同的建模方法进行研究，因此区分研究对象是平稳时间序列还是非平稳时间序列是时间序列分析的首要步骤。

时间序列分析理论中有两种平稳性定义，即所谓严平稳性（Strictlystationary）和弱平稳性（Weaklystationary）。

2.2.1 严平稳性

严平稳性也称强平稳性（Stronglystationary），是一种条件比较苛刻的平稳性定义，它认为只有当序列所有的统计性质都不会随着时间的推移而发生变化时，该序列才能被认为平稳。而我们知道，随机变量族的统计性质完全由他们的联合概率分布族决定，所以严平稳时间序列的定义如下。

定义：设 $\{X_t\}$ 为一时间序列，对任意正整数 m，任取 $t_1,t_2,\cdots,t_m \in T$，对任意整数 τ，有：

$$F_{t_1,t_2,\cdots,t_m}(X_1,X_2,\cdots,X_m)=F_{t_1+\tau,t_2+\tau,\cdots,t_m+\tau}(X_1,X_2,\cdots,X_m)$$

则称时间序列 $\{X_t\}$ 为严平稳时间序列。

在实践中要获得随机序列的联合分布是一件非常困难的事，而且即使知道随机序列的联合分布，计算和应用起来也非常不便。所以严平稳时间序列通常只具有理论意义，在实践中用得更多的是条件比较宽松的弱平稳时间序列。

2.2.2 弱平稳性

弱平稳性也称协方差平稳性（Covariance Stationary）、二阶平稳性（Second-order Stationary）或宽平稳性（Wide-sense Stationary），它是在时间序列二阶矩基础上定义的平稳性。简单来说，弱平稳时间序列的一阶矩和二阶矩不随时间的变化而改变。

弱平稳性（Weakstationary）是使用序列的特征统计量来定义的一种平稳性。它认为序列的统计性质主要由它的低阶矩决定，所以只要保证序列低阶（二阶）矩平稳，就能保证序列的主要性质近似稳定。

定义：如果 $\{X_t\}$ 满足以下三个条件：

（1）任取 $t, j \in T$，有 $\mathrm{Var}(X_t) = \mathrm{Var}(X_{t-j}) = \sigma^2$

（2）任取 $t, j \in T$，$E(X_t) = E(X_{t-j}) = \mu$，$\mu$ 为常数

（3）任取 $t, j, s \in T$，有 $\gamma(t, t-s) = \gamma(t-j, t-s-j) = \gamma(s)$

则称 $\{X_t\}$ 为宽平稳时间序列，宽平稳也称为弱平稳或二阶平稳（Second-order Stationary）。

弱平稳定义中，（1）和（2）表明弱平稳时间序列具有有限的常数均值和方差，（3）表明弱平稳时间序列的自协方差只与时滞 s 有关，而与时间的起始位置无关。因此可以将自协方差函数由二维函数 $\gamma(t,s)$ 简化为一维函数 $\gamma(s-t)$。概括来说，弱平稳时间序列的一阶矩和二阶矩都是不随时间变化而改变的常数。

由于平稳时间序列的自相关系数是时滞 s 的函数，因此通常也称 ρ_s 为自相关函数（Auto-Correlation Function, ACF），ρ_s 对时滞 s 作图通常称为自相关图（Correlogram）。

$$\rho_s = \frac{\gamma_s}{\gamma_0} = \frac{\gamma_s}{\sigma^2}$$

容易验证，和相关系数一样，自相关系数具有如下三个性质：

（1）规范性

$$\rho_0 = 1 \text{ 且 } |\rho_k| \leqslant 1, \quad \forall k$$

（2）对称性

$$\rho_k = \rho_{-k}$$

（3）非负定性

对任意正整数 m，相关阵 Γ_m 为对称非负定阵。

$$\Gamma_m = \begin{pmatrix} \rho_0 & \rho_1 & \cdots & \rho_{m-1} \\ \rho_1 & \rho_0 & \cdots & \rho_{m-2} \\ \vdots & \vdots & \rho_0 & \vdots \\ \rho_{m-1} & \rho_{m-2} & \cdots & \rho_0 \end{pmatrix}$$

值得注意的是 ρ_k 除了具有这三个性质外，它还具有一个特别的性质：非唯一性。

一个平稳时间序列一定唯一决定了它的自相关函数，但一个自相关函数未必唯一对应着一个平稳时间序列。在本书所涉及的时间序列建模理论中，只考虑平稳性即可，因此本书中后续涉及的平稳性都指弱平稳性。

一般来说，满足严平稳的序列也具有弱平稳性，但严平稳却并不能全部涵盖弱平稳。例如，如果一个严平稳时间序列不存在二阶矩或一阶矩（如柯西分布），则它就不满足弱平稳性。

2.2.3 时序图检验

所谓时序图就是一个平面二维坐标图，通常横轴表示时间，纵轴表示序列取值，时序图可以直观地帮助我们掌握时间序列的一些基本分布特征。

根据平稳时间序列均值、方差为常数的性质，平稳序列的时序图应该显示出该序列始终在一个常数值附近随机波动，而且波动范围有界的特点。如果观察序列的时序图显示出该序列有明显的趋势性或周期性，那它通常不是平稳序列。根据这个性质，很多非平稳序列通过查看它的时序图可以立刻被识别出来。

绘制 2017 年 12 月 5 日到 12 月 29 日的上证综指收盘价的时间序列，可得到如图 2-1 所示的趋势图。

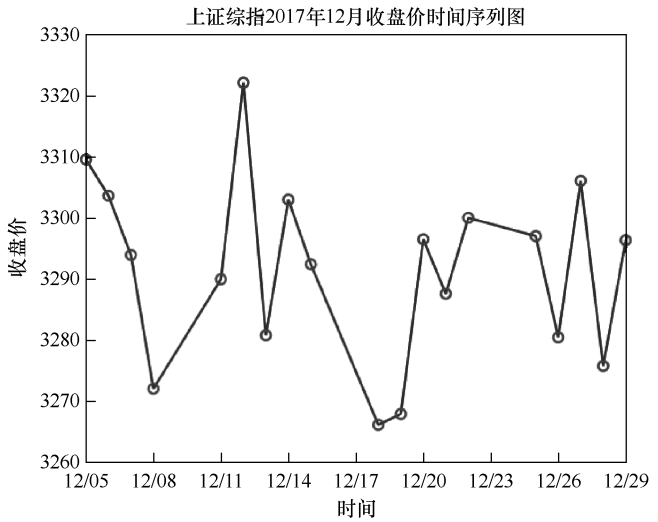


图 2-1　2017 年上证综指收盘价的时间序列

可以看出 12 月收盘价的时间序列围绕一个常数值上下波动，很可能是一个平稳的序列。而 2017 年 5—6 月上证综指的时序图呈现明显的向上趋势，如图 2-2 所示，该时间段内的时间序列明显不是平稳的。更严谨的平稳性检验将在后面的章节中介绍。

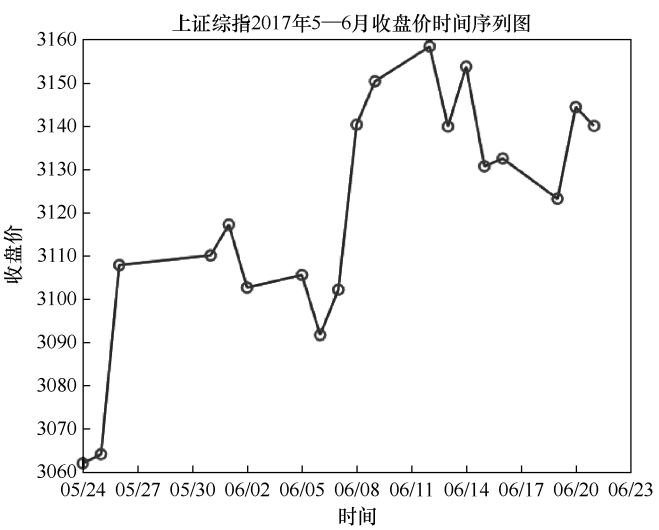


图 2-2　2017 年上证综指 5—6 月收盘价时间序列图

2.3 时间序列的相关性

序列相关性，又称自相关（Autocorrelation），是指总体回归模型的随机误差项之间存在相关关系。在计量经济学中指对于不同的样本值，随机干扰之间不再是完全相互独立的，而是存在某种相关性。

在回归模型的古典假定中，假设随机误差项是无自相关的，即在不同观测点之间是不相关的。如果该假定不能满足，就称存在自相关，即不同观测点上的误差项彼此相关。

自相关的程度可用自相关系数去表示，根据自相关系数的符号可以判断自相关的状态，如果小于 0，则 X_t 与 X_{t-1} 为负相关；如果大于 0，则 X_t 与 X_{t-1} 为正关；如果等于 0，则 X_t 与 X_{t-1} 不相关。

2.3.1 自相关的分类

我们按自相关表现形式分类，可以将其分为以下两种类型。

（1）如果误差项只与其滞后一期的值相关，则称误差项存在一阶自相关。即：

$$\varepsilon_t = f(\varepsilon_{t-1}) + v_t, t = 1,2,\cdots,T$$

（2）如果误差项与其滞后若干期（大于 1 期）的值相关，则误差项存在高阶自相关。即：

$$\varepsilon_t = f(\varepsilon_{t-1},\varepsilon_{t-2},\cdots) + v_t, t = 1,2,\cdots,T$$

计量经济学中最常见的自相关形式为一阶线性自回归形式，即：$\varepsilon_t = \alpha_1\varepsilon_{t-1} + v_t$。其中，$\alpha_1$ 称为自回归系数；v_t 是满足标准 OLS 假定的随机误差项。

$$E(v_t) = 0, t = 1,2,\cdots,T$$

$$\mathrm{Var}(v_t) = \sigma_v^2, t = 1,2,\cdots,T$$

$$\mathrm{Cov}(v_i, v_j) = 0, i \neq j, t = 1,2,\cdots,T$$

$$\mathrm{Cov}(\varepsilon_{t-1}, v_t) = 0, t = 1,2,\cdots,T$$

根据最小二乘原理和相关系数的定义，可以得到：

$$\hat{\alpha}_1=\frac{\sum_{t=2}^{T}\varepsilon_t\varepsilon_{t-1}}{\sum_{t=2}^{T}\varepsilon_{t-1}^2}\approx\hat{\rho}=\frac{\sum_{t=2}^{T}\varepsilon_t\varepsilon_{t-1}}{\sqrt{\sum_{t=2}^{T}\varepsilon_t^2}\sqrt{\sum_{t=2}^{T}\varepsilon_{t-1}^2}}\Rightarrow\rho\approx\alpha_1$$

即在大样本条件下，一阶自回归系数等于这二个变量的相关系数。由此，误差项的一阶线性自回归形式可写为：

$$\varepsilon_t=\rho\varepsilon_{t-1}+v_t,-1\leqslant\rho\leqslant1$$

如相关系数大于 0，则称误差项存在正自相关；如相关系数小于 0，则称误差项存在负自相关。

注意：自相关不是指两个或两个以上的变量之间的相关关系，而是指一个变量前后期数值之间存在的相关关系。

2.3.2 自相关的来源

1．经济变量固有的惯性

大多数经济时间序列数据都有一个明显的特点——惯性，表现为滞后值对本期值具有影响。例如，GDP、价格指数、生产、就业与失业等时间序列都呈周期性，如周期中的复苏阶段，大多数经济序列均呈上升势，序列在每一时刻的值都高于前一时刻的值，似乎有一种内在的动力在驱使这一势头继续下去，直至某些情况（如利率或课税的升高）出现才把它拖慢下来。

2．模型设定的偏误

所谓模型设定偏误（Specification Error）是指所设定的模型“不正确”。主要表现为在模型中丢掉了重要的解释变量或模型函数形式有偏误。

例如，本来应该估计的模型为：

$$Y_t=\beta_0+\beta_1X_{1t}+\beta_2X_{2t}+\beta_3X_{3t}+u_t$$

但在模型设定中做了下述回归：

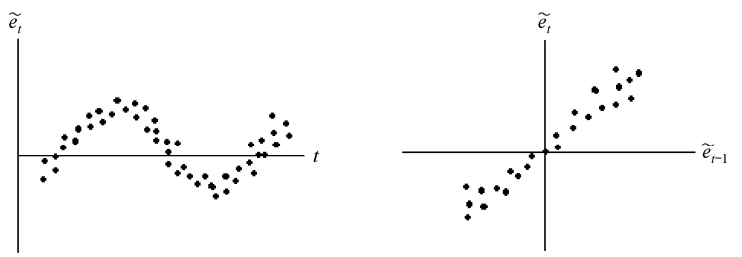
$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + v_t$$

因此，$v_t = \beta_3 X_{3t} + u_t$。如果 X_3 确实影响 Y_t，则随机误差项中有一个重要的系统性影响因素，使其呈序列相关性。

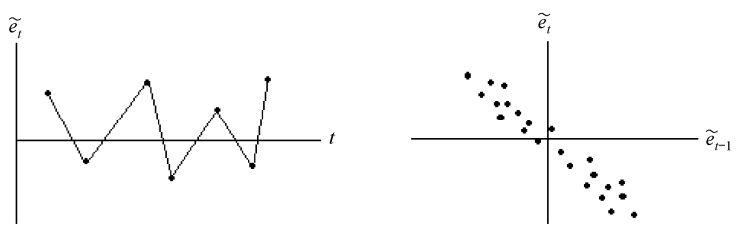
2.3.3 自相关的检验

1．图示法

由于残差项 e_i 可以作为随机误差项 ε_i 的近似估计，因此如果 ε_i 存在序列相关，必然由残差项 e_i 反映出来。因此可利用 e_i 的变化图来判断随机误差项 ε_i 的序列相关性，如图 2-3 所示。



正序列相关（正自相关）



负序列相关（负自相关）

图 2-3　残差相关性示意图

2．德宾−沃森（Durbin-Watson）检验法

D-W 检验是德宾（J.Durbin）和沃森（G.S.Watson）于 1951 年提出的一种检验序列自相关的方法。

该方法的使用条件是：

① 解释变量 X 非随机，或者在重复抽样中被固定；

② 随机误差项 ε_i 为一阶自回归形式：$\varepsilon_i = \rho\varepsilon_{i-1} + \nu_i$；

③ 回归模型中不应出现滞后被解释变量作为解释变量，即不应该出现下列形式：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \gamma Y_{i-1} + \varepsilon_i$$

④ 回归含有截距项；

⑤ 没有缺失数据。

该方法的检验假设是：

$\mathrm{H}_0 : \rho = 0$，即随机误差项不存在一阶序列相关

$\mathrm{H}_1 : \rho \neq 0$，即随机误差项存在一阶序列相关

构造的统计量是：

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

该统计量的分布与给定样本中的 X 值有复杂的关系，因此其精确的分布很难得到。但是，Durbin 和 Watson 成功地导出了临界值的下限 d_L 和上限 d_U，且这些上下限只与样本的容量 n 和解释变量的个数 k 有关，而与解释变量 X 的取值无关。

进行 DW 检验的步骤如下：

① 计算 D.W.统计量的值；

② 根据样本容量 n 和解释变量数目 k 查阅 D.W.分布表，得到临界值 d_L 和 d_U；

③ 按照表 2-1 的准则考察计算得到的 D.W.值，以判断随机误差项是否存在一阶自相关。

表 2-1　DW 检验决策规则

DW 统计量	相应准则
$0 \leqslant DW \leqslant d_L$	误差项 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 间存在正相关
$d_L < DW \leqslant d_U$	不能判定是否有自相关
$d_U < DW < 4-d_U$	误差项 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 间无自相关
$4-d_U \leqslant DW < 4-d_L$	不能判定是否有自相关
$4-d_L \leqslant DW \leqslant 4$	误差项 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 间存在负相关

用坐标图可以更直观地表示 DW 检验规则，如图 2-4 所示。

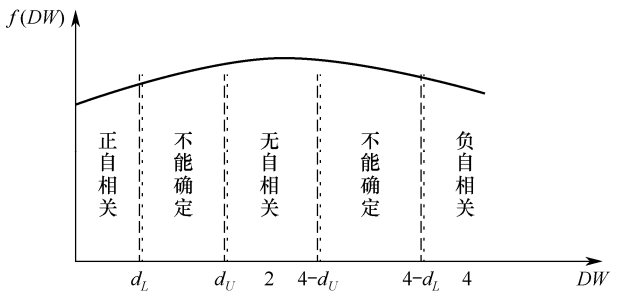


图 2-4　DW 检验规则

容易证明，当 D.W.值在 2 左右时，模型不存在一阶自相关。

如果存在完全一阶正相关，即ρ=1，则 D.W.≈0；

如果存在完全一阶负相关，即ρ= −1，则 D.W.≈4；

如果完全不相关，即ρ=*0*，则 D.W.≈2。

注意：

① 从判断准则可以看到，存在一个不能确定的 D.W.值区域，这是这种检验方法的一大缺陷。

② D.W.检验虽然只能检验一阶自相关，但在实际计量经济学问题中，一阶自相关是出现最多的一类序列相关。

③ 经验表明，如果不存在一阶自相关，一般也不存在高阶序列相关。

所以在实际应用中，对于序列相关问题一般只进行 D.W.检验。

3．LM 检验（或 BG 检验）

此方法不仅适用于一阶自相关检验，也适用于高阶自相关的检验。

检验步骤：

（1）用 OLS 对回归模型进行，得到残差序列 e_t；

（2）运用残差序列和样本观测值中的解释变量，建立如下辅助回归模型并进行 OLS 估计，得到样本可决系数 R^2；

$$e_t = \hat{\rho}_1 e_{t-1} + \hat{\rho}_2 e_{t-2} + \cdots + \hat{\rho}_n e_{t-n} + \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + v_t$$

LM 检验的检验假设为：

$$\mathrm{H}_0 : \rho_1 = \rho_2 = \cdots = \rho_n = 0$$

（3）构造 LM 统计量：$LM = TR^2 \overset{近似}{\sim} \chi^2(n)$

（4）查 χ^2 分布表，求得临界值：$\chi_\alpha^2(n)$

若 $LM = TR^2 > \chi_\alpha^2(n)$，则拒绝原假设，说明随机误差存在序列相关性，反之不能拒绝原假设。

2.3.4 自相关的解决方法

如果随机误差项被检验证明存在序列相关性，首先应分析产生自相关的原因，如果是由于模型设定偏误，则应修改模型的数学形式。

怎样查明自相关是由模型设定偏误引起的？一种方法是用残差对解释变量进行较高次幂回归，然后对新残差作 DW 检验，如果此时自相关消失，则说明模型设定存在偏误。

如果模型产生自相关的原因是模型中省略了重要解释变量，则解决方法就是找出被省略了的解释变量，将其作为解释变量列入模型。怎样查明此种自相关呢？一种方法是用残差 e_t 对那些可能影响被解释变量而未被列入模型的解释变量进行回归，并作显著性检验，从而确定该解释变量的重要性。

只有当以上两种引起自相关的原因都消除以后，才能认为随机误差项“真正”存在自相关，此时需要对原模型进行变换，使变换以后的模型的随机误差项的自相关得以消除，进而利用普通最小二乘法估计回归参数。

最常用的方法是广义最小二乘法（GLS：Generalized Least Squares），这种方法对原模型进行适当变换以消除误差项的自相关，进而利用 OLS 来估计回归参数，相应的回归参数估计结果称为广义最小二乘估计量。

2.4 时间序列的运算

人们常常通过建立模型来描述现象、事物随时间推移的变化规律性，而常见的模型一般都是某种运算的结果。

2.4.1 线性运算和延迟运算

1．线性运算

设 $\{X_t\}$，$\{Y_t\}$ 是两个时间序列，a，b 是两个常数，形如 $\{aX_t \pm bY_t\}$ 的时间序列可以看作由 $\{X_t\}$、$\{Y_t\}$ 经过线性运算得到。

注意：不相关平稳序列的线性运算能保持平稳性，其他的线性运算不一定能保持平稳性。

2．延迟运算

假设已知时间序列 $\{X_t\}$ 和 $\{Y_t\}$ 有如下关系 $X_t = Y_{t-1}$，这一关系式也可以用滞后算子（也称延迟算子）L 表示为：

$$X_t = LY_t$$

这意味着在各时间点上，时间序列 $\{X_t\}$ 和 $\{Y_t\}$ 的对应关系如表 2-2 所示。

表 2-2　$X_t=LY_t$ 关系下时间序列$\{X_t\}$和$\{Y_t\}$的对应关系

时期 / 序列	1	2	3	4	5	6	7	8	9	10	…
X_t	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	…
Y_t	—	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	…

因此，滞后算子 L 的作用就是将时间序列逐项推后一期。关于滞后算子，有下面的 2 种关系式成立：$LX_t = X_{t-1}$，$L^j X_t = X_{t-j}$。

其中，j 为正数。称 L 为一步延迟算子，L^2 为二步延迟算子，…，L^j 为 j 步延迟算子。

当 $j=0$ 时，$L^0 X_t = X_t$，当 $j=-i$ 时，$L^{-i} X_t = X_{t+i}$。

另外，滞后算子具有如下性质：

（1）对常数施加滞后算子仍为常数，即 $L\mathrm{c}=\mathrm{c}$，其中，c 表示常数。

（2）滞后算子适用分配率，即 $\left(L^i + L^j\right) X_t = L^i X_t + L^j X_t = X_{t-i} + X_{t-j}$。

（3）滞后算子适用结合率，即 $L^i L^j X_t = L^i \left(L^j X_t\right) = L^i X_{t-j} = X_{t-i-j}$。

还可以通过线性运算，构造滞后算子多项式来对时间序列进行更加复杂的运算。典型的 p 阶滞后算子多项式为：$A(L) = a_0 + a_1 L + a_2 L^2 + \cdots + a_p L^p$，则滞后算子多项式 $A(L)$ 施加于时间序列时有：

$$\begin{aligned} A(L)X_t &= \left(a_0 + a_1 L + a_2 L^2 + \cdots + a_p L^p\right) X_t \\ &= a_0 X_t + a_1 L X_t + a_2 L^2 X_t + \cdots + a_p L^p X_t \\ &= a_0 X_t + a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} \end{aligned}$$

特别地，当 $p \to \infty$，且滞后算子多项式中的系数 $a_j = a^j$ 时，则构成特殊的无限期滞后算子多项式：

$$1 + aL + a^2 L^2 + \cdots = \sum_{j=0}^{\infty} a^j L^j$$

若 $|a| < 1$，则上式可记为：

$$1 + aL + a^2 L^2 + a^3 L^3 + \cdots = \sum_{j=0}^{\infty} a^j L^j = \frac{1}{1-aL}$$

此外还有线性延迟混合运算，设 $Y_t = a_0 X_t + a_1 X_{t-1} + \cdots + a_p X_{t-p}$，称 Y_t 为 $\{X_t\}$ 的 p 阶滑动平均。

2.4.2 差分算子

1. 一阶差分

时间序列分析过程中，经常会用到差分运算。对于时间序列$\{X_t\}$，差分运算可以表示为：

$$\Delta X_t = X_t - X_{t-1} = (1-L)X_t$$

其中，Δ为差分算子。对于差分后的时间序列$\{\Delta X_t\}$来说，它与原序列$\{X_t\}$之间的关系如表 2-3 所示。

表 2-3 时间序列$\{X_t\}$和差分序列$\{\Delta X_t\}$的对应关系

时期 / 序列	1	2	3	4	5	6	7	8	…
X_t	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	…
ΔX_t	—	X_2-X_1	X_3-X_2	X_4-X_3	X_5-X_4	X_6-X_5	X_7-X_6	X_8-X_7	…

2. 高阶差分

差分后的序列$\{\Delta X_t\}$仍然可以再次进行差分，即$\{\Delta\Delta X_t\}$，这相对于原序列$\{X_t\}$来说是做了二次差分，方便起见差分后再差分记作Δ^2，并称之为二阶差分。二阶差分序列$\{\Delta^2 X_t\}$与原序列$\{X_t\}$之间的关系可以表示为：

$$\begin{aligned}\Delta^2 X_t &= \Delta\Delta X_t = \Delta(X_t - X_{t-1}) = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\ &= X_t - 2X_{t-1} + X_{t-2} = (1-2L+L^2)X_t\end{aligned}$$

或

$$\begin{aligned}\Delta^2 X_t &= \Delta X_t - \Delta X_{t-1} = (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\ &= X_t - 2X_{t-1} + X_{t-2} = (1-2L+L^2)X_t\end{aligned}$$

因此，可以认为差分运算没有顺序，如果考虑更高阶的p阶差分，有：

$$\Delta^p = \Delta\Delta^{p-1} = \Delta\Delta\Delta^{p-2} = \cdots = \underbrace{\Delta\cdots\Delta}_{p} = \cdots = \Delta^{p-2}\Delta\Delta = \Delta^{p-1}\Delta$$

3．s 步差分

另外，差分运算还可以运用于更多间隔的时间点之间。考虑：

$$\Delta_s X_t = X_t - X_{t-s} = (1-L^s)X_t$$

其中，Δ_s 称为 s 步差分。

s 步差分的典型用处是测算季度或月度时间序列数据的同比变化，因此也称季节性差分。例如，一个季度时间序列每年有 4 个季度的数据，如果希望了解每年数据与上年同期相比的变化情况，则可以用四步差分序列来刻画。时间序列与其四步差分序列之间的对应关系如表 2-4 所示。

表 2-4　时间序列$\{X_t\}$和四步差分序列$\{\Delta_4 X_t\}$的对应关系

时期 / 序列	1	2	3	4	5	6	7	8	…
X_t	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	…
$\Delta_4 X_t$	—	—	—	—	X_5-X_1	X_6-X_2	X_7-X_3	X_8-X_4	…

4．联合 p 阶差分和 s 步差分

对于一个时间序列来说，除了可以单独进行 p 阶差分和 s 步差分，有时还需要两种差分运算联合进行。例如，在后面的时间序列建模过程中经常看到一阶四步差分 $\Delta\Delta_4$ 或一阶十二步差分 $\Delta\Delta_{12}$ 的情况。值得注意的是，p 阶差分和 s 步差分联合运算时，没有先后顺序之分，即对于时间序列 $\{X_t\}$，有 $\Delta^p\Delta_s X_t = \Delta_s\Delta^p X_t$

2.5　白噪声

2.5.1　纯随机序列

拿到一个观察值序列之后，首先判断它的平稳性。通过平稳性检验，序列可以分为平稳序列和非平稳序列两大类。

对于非平稳序列，由于它不具有二阶矩平稳的性质，所以对它的统计分析要复杂

一些，通常要进行进一步的检验、变换或处理之后，才能确定适当的拟合模型。

如果序列平稳，情况就简单多了，我们有一套非常成熟的平稳序列建模方法。但是，并不是所有的平稳序列都值得建模。只有那些序列值之间具有密切的相关关系，历史数据对未来的发展有一定影响的序列，才值得我们花时间去挖掘历史数据中的有效信息，用来预测序列未来的发展。

如果序列值彼此之间没有任何相关性，那就意味着该序列是一个没有记忆的序列，过去的行为对将来的发展没有丝毫影响，这种序列我们称之为纯随机序列。从统计分析的角度而言，纯随机序列是没有任何分析价值的序列。

为了确定平稳序列还值不值得继续分析下去，我们需要对平稳序列进行纯随机性检验。

2.5.2 白噪声过程

时间序列分析的主要内容是对时间序列建模，而时间序列模型中的随机性往往是通过白噪声（Whitenoise）过程来引入的。白噪声过程可谓是构建时间序列模型的基石。

定义：若$\{\varepsilon_t\}$满足零均值、同方差和非自相关，即：

$$E(\varepsilon_t)=0$$
$$\mathrm{Var}(\varepsilon_t)=\sigma_\varepsilon^2$$
$$E(\varepsilon_t\varepsilon_{t-s})=0$$

对所有的t和$s\neq t$成立，则称$\{\varepsilon_t\}$为白噪声过程，通常可记作$\varepsilon_t \sim \mathrm{WN}(0,\sigma_\varepsilon^2)$。特别地，如果$\varepsilon_t$服从正态分布，则称$\{\varepsilon_t\}$为正态白噪声过程或高斯白噪声过程。

从白噪声过程的定义可以看出，它满足平稳性条件，因此白噪声过程是一种特殊的平稳性过程。白噪声过程的显著特点是高度的随机性，也称纯随机性，即各时刻随机变量之间互不相关，因此就没有必要构建时间序列模型再去研究其相关关系。一般情况下，对一个时间序列进行研究，当各时刻随机变量之间所有的相关关系都已经通过建模被识别后，剩下的无须进一步研究的部分通常都是白噪声过程，也就是在这种意义上，白噪声过程成了时间序列模型的基本构件。

利用 MATLAB 根据定义可以很容易得到白噪声的序列图，比如产生均值为 0，方差为 1 的白噪声序列图，如图 2-5 所示。主要实现方法是通过 mvnrnd(0,1,1000)命令产生均值为 0，方差为 1 的 1000 个随机数。

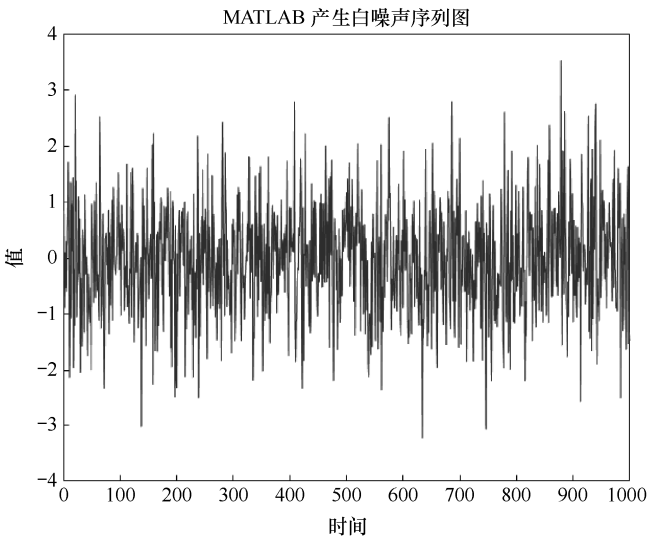


图 2-5　白噪声序列图

白噪声序列虽然很简单，但它在我们进行时间序列分析中所起的作用却非常大，它的两个重要性质我们在后面的分析过程中要经常用到。

1．纯随机性

由于白噪声序列具有如下性质：

$$\gamma(k)=0,\quad \forall k\neq 0$$

这说明白噪声序列的各项之间没有任何相关关系，这种“没有记忆”的序列就是我们说的纯随机序列。纯随机序列各项之间没有任何关联，序列在进行完全无序的随机波动。一旦某个随机事件呈现出纯随机运动的特征，我们就认为该随机事件没有包含任何值得提取的有用信息，我们就应该终止分析了。

如果序列值之间呈现出某种显著的相关关系：

$$\gamma(k)\neq 0,\quad \exists k\neq 0$$

就说明该序列不是纯随机序列，该序列间隔期的序列值之间存在一定程度的相互影响关系，这种相互影响的关系，统计上称为相关信息。我们分析的目的就是要想方设法把这种相关信息从观察值序列中提取出来。一旦观察值序列中蕴含的相关信息被我们充分提取出来了，那么剩下的残差序列就应该呈现出纯随机的性质了，所以纯随机性还是我们判断相关信息是否提取充分的一个判别标准。

2．方差齐性

所谓方差齐性，就是指序列中每个变量的方差都相等，即：

$$DX_t = \sigma^2$$

如果序列不满足方差齐性，我们就称该序列具有异方差性质。

在时间序列分析中，方差齐性是一个非常重要的限制条件。因为根据马尔可夫定理，只有方差齐性假设成立时，我们用最小二乘法得到的未知参数估计值才是准确的、有效的。如果假设不成立，那么最小二乘估计值就不是方差最小线性无偏估计，拟合模型的预测精度会受到很大影响。

所以我们在进行模型拟合时，检验内容之一就是要检验拟合模型的残差是否满足方差齐性假定。如果不满足，那就说明残差序列还不是白噪声序列，即拟合模型没有充分提取随机序列中的相关信息，这时拟合模型的精度是值得怀疑的。在这种情况下，我们通常需要使用适当的条件异方差模型来拟合该序列。

2.6 小结

本章介绍了时间序列的基本概念。主要包括时间序列的统计概念，详细讲述了其联合分布函数、边际分布、概率分布及常见的特征统计量，例如均值、方差、自协方差函数等，这些基本概念是贯穿整本书中的基础知识点。其次介绍了时间序列建模的重要性质：平稳性，平稳时间序列行为不随时间改变，这是确定模型的基础性质。再次介绍了时间序列对的相关性，相关性说明了时间序列数据之间的相关关系。在 2.4 节介绍了时间序列的运算，包括线性运算、延迟运算及差分运算，这些运算在后面的章节中对时间序列数据的处理起到了至关重要的作用。最后一节介绍了时间序列数据

中很重要的一项——白噪声，这是有关拟合模型残差的重要假设。本章主要是讲述基础知识，为后续的建模奠定基础。

参考文献

[1] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[2] Ruey S. Tsay, 王远林，王辉，等. 金融时间序列分析（第三版）. 北京：人民邮电出版社，2012.

[3] Jonathan D.Cryer, Kung-Sik Chan. Time Series Analysis With Applications in R (Second Edition): New York, Springer, 2008.

[4] Hamilton. James, 夏晓华. 时间序列分析. 北京：中国人民大学出版社，2015.

[5] 王黎明，王连，杨楠. 应用时间序列分析. 上海：复旦大学出版社，2009.

[6] 黄红梅. 应用时间序列分析. 北京：清华大学出版社，2016.

3 自回归模型——AR 模型

本章将介绍时间序列最基本的模型——自回归模型。

3.1 AR 模型的定义

自回归模型（Autoregressive Model，AR 模型）是一种从线性回归分析中发展而来的处理时间序列的方法。该方法是用自身做回归变量的过程，即利用前期若干时刻的随机变量的线性组合来描述以后某时刻随机变量的线性回归过程。与其他线性回归相比，自回归并不是用 x 预测 y，而是用 x 预测 x（自己）。自回归模型被广泛运用在经济学、信息学、自然现象等的预测上。

定义：设时间序列$\{X_t\}$，满足：

$$X_t=a_0+a_1X_{t-1}+\cdots+a_pX_{t-p}+\varepsilon_t$$

式中$\{\varepsilon_t\}$是白噪声序列，$a_0,a_1,\cdots,a_p$是p+1 个实数，称此模型为 p 阶自回归模型，记为 AR（p）模型，称适合此模型的$\{X_t\}$为 AR（p）序列。

当 a_0=0 时，称为中心化的 AR（p）模型，后文多讨论中心化的模型。

对于$\forall s<t, E(X_s\varepsilon_t)=0$，要求前面的时间序列与后面的白噪声不相关，此条件称为合理性条件。

一般的，可以定义在 AR（p）中的系数多项式$a(u)=1-a_0-a_1u-\cdots-a_pu^p$为 AR（$p$）模型的自回归系数多项式。

令$\partial(L)=1-a_1L-a_2L^2-\cdots-a_pL^p$，则 AR（$p$）模型的算子表达式可表示为：

$$\partial(L)X_t=\varepsilon_t$$

下面我们讨论 AR 模型的求解，先讨论 AR（1）模型 $X_t=aX_{t-1}+\varepsilon_t$，在 $|a|<1$ 时模型的平稳解。

将 $X_t=aX_{t-1}+\varepsilon_t$ 移项得：

$$X_t-aX_{t-1}=\varepsilon_t$$

$$(1-aL)X_t=\varepsilon_t$$

所以可得：

$$X_t=\frac{1}{1-aL}\varepsilon_t=\sum_{j=0}^{+\infty}(aL)^j\varepsilon_t=\sum_{j=0}^{+\infty}a^j\varepsilon_{t-j}$$

因 $|a|<1$，故 $\sum_{j=0}^{+\infty}|a^j|<+\infty$，解为平稳解。

一般模型 AR（p）的解可通过类似方式求得为：

$$a(u)=1-a_0-a_1u-\cdots-a_pu^p,\partial(L)X_t=\varepsilon_t$$

$$X_t=\frac{1}{a(L)}\varepsilon_t=\sum_{j=0}^{\infty}\phi_jL^j\varepsilon_t=\sum_{j=0}^{\infty}\phi_j\varepsilon_{t-j},\phi_0=1$$

3.2 AR 模型的平稳性

3.2.1 AR 模型的平稳条件

AR 模型是常用的平稳序列的拟合模型之一，但并非所有的 AR 模型都是平稳的，因此在使用 AR 模型之前需要进行平稳性的判断。根据自回归系数多项式，定义平稳性条件如下：若 $a(u)=0$ 的根都在单位圆外时，称此为平稳的 AR（p）模型，否则为非平稳的 AR（p）模型，或者广义的 AR（p）模型。

即平稳条件：$\sum_{j=0}^{\infty}\left|\phi_j\right|<+\infty$ 或 $\sum_{j=0}^{\infty}\phi_j^2<+\infty$ 满足时，

$$X_t=\sum_{j=0}^{\infty}\phi_j\varepsilon_{t-j},\phi_0=1 \text{ 平稳。}$$

此外还有平稳域判别法：称$\{a \mid a(u)=0\}$的根在单位圆外，$a^T=(a_1,a_2,\cdots,a_p)\in \mathrm{R}^P$为 AR（$p$）模型的平稳域。

3.2.2 AR（1）模型的平稳域

对于中心化平稳 AR（1）模型 $X_t=aX_{t-1}+\varepsilon_t$，令其系数多项式等于 0，即 $1-au=0$，则 $u=\frac{1}{a}$，所以平稳域是 $\{a\left|0<\right|a\left|<1\right\}$。

二阶自回归模型 $X_t=a_1X_{t-1}+a_2X_{t-2}+\varepsilon_t$ 中，方程 $1-a_1u-a_2u^2=0$ 的两根分别为 u_1，u_2，则：

$$u_1,u_2=\frac{a_1\pm\sqrt{a_1^2+4a_2}}{-2a_2}$$

为了满足平稳条件，要求根的绝对值大于 1，因此要满足：

$$a_1+a_2<1,a_2-a_1<1,\left|a_2\right|<1$$

对高阶自回模型 AR（p）来说，多数情况下没有必要直接计算其特征方程的特征根，但有一些有用的规则可以用来检验高阶自回归模型的稳定性。

AR（p）模型稳定的必要条件是：

$$a_1+a_2+\cdots+a_p<1$$

由于 $a_i(i=1,\cdots,p)$可正可负，AR（p）模型稳定的充分条件是：

$$\left|a_1\right|+\left|a_2\right|+\cdots+\left|a_p\right|<1$$

AR（2）过程平稳参数区域如图 3-1 所示。

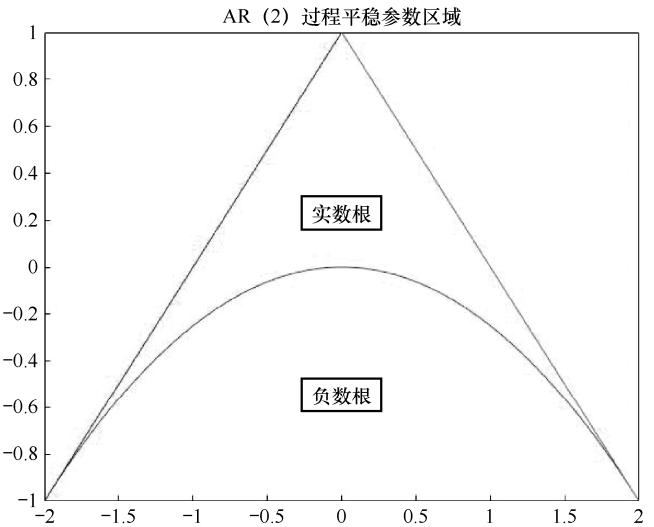


图 3-1　AR（2）过程平稳参数区域

3.3　AR 模型的统计性质

3.3.1　均值

如果 AR（p）模型满足平稳性条件，则有：

$$E(X_t)=E(a_0+a_1X_{t-1}+\cdots+a_pX_{t-p}+\varepsilon_t)$$

根据平稳序列均值为常数，且$\{\varepsilon_t\}$为白噪声序列，有

$$E(X_t)=\mu, E(\varepsilon_t)=0, \forall t\in T$$

由此可得：

$$\mu=\frac{a_0}{1-a_1-\cdots-a_p}$$

3.3.2　方差

将平稳的 AR（p）模型表示成如下的传递形式：

$$X_t=\frac{\varepsilon_t}{\partial(L)}=\sum_{i=1}^{P}\frac{k_i}{1-\lambda_i L}\varepsilon_t=\sum_{i=1}^{p}\sum_{j=0}^{\infty}k_i(\lambda_i L)^j\varepsilon_t=\sum_{j=0}^{\infty}\sum_{i=1}^{\infty}k_i\lambda_i^j\varepsilon_{t-j}\triangleq\sum_{j=0}^{\infty}G_j\varepsilon_{t-j}$$

其中系数 $\{G_j, j=1,2,\cdots\}$ 被称为 Green 函数。

由平稳 AR 模型的传递形式：

$$X_t=\sum_{j=0}^{\infty}G_j\varepsilon_{t-j}$$

两边求方差得：

$$\mathrm{Var}(X_t)=\sum_{j=0}^{\infty}G_j^2\sigma_\varepsilon^2$$

3.3.3 自协方差函数

在平稳 AR（p）模型两边同乘 $X_{t-k}, \forall k\geqslant 1$，再求期望：

$$E(X_tX_{t-k})=a_1E(X_{t-1}X_{t-k})+\cdots+a_pE(X_{t-p}X_{t-k})+E(\varepsilon_tX_{t-k})$$

根据 $E(\varepsilon_tX_{t-k})=0, \forall k\geqslant 1$ 可以得到自协方差函数的递推公式：

$$\gamma_k=a_1\gamma_{k-1}+\cdots+a_p\gamma_{k-p}$$

3.3.4 自相关系数

自相关系数的定义是：$\rho_k=\frac{\gamma_k}{\gamma_0}$，特别地：$\rho_0=1, \rho_1=\frac{a_1}{1-a_2}$

平稳 AR（p）模型的自相关系数递推公式：

$$\rho_k=a_1\rho_{k-1}+\cdots+a_p\rho_{k-p}$$

上述方程称为 Yule-Walker 方程,

注意：在 AR（1）模型中，即使 X_{t-2} 没有直接出现在模型中，X_{t-2} 和 X_t 也是相关的，因为 $X_{t-1}=a_1X_{t-2}+\varepsilon_{t-1}$。

所以，X_{t-2} 是通过 X_{t-1} 与 X 相关的，这种间接相关出现在所有 AR 模型中。

X_{t-2} 与 X_t 的自相关系数 ρ_2 等于 X 与 X_{t-1} 的自相关系数 ρ_1 乘 X_{t-1} 与 X_t 的自相关系数 ρ_1，即 $\rho_2=\rho_1^2$。

平稳 AR（p）模型的自相关系数有拖尾性。拖尾性说明 X_t 之前的每一个序列值 X_{t-1}，X_{t-2}，…都会对 X_t 构成影响，但因为自相关系数呈负指数衰减，所以间隔较远的序列值对现时值的影响很小，具有所谓的“短期相关性”。

3.3.5 偏自相关函数

自相关函数 ACF（k）给出了 X_t 与 X_{t-k} 的总体相关性，但总体相关性可能掩盖了变量间完全不同的相关关系。

例如，在 AR（1）中，X_t 与 X_{t-2} 间有相关性可能主要是由于它们各自与 X_{t-1} 间的相关性带来的：

$$\rho_2=\rho_1^2=E(X_tX_{t-1})E(X_{t-1}X_{t-2})$$

即自相关函数中包含了这种所有的“间接”相关。

与之相反，X_t 与 X_{t-k} 间的偏自相关函数（Partial Autocorrelation，PACF）则是消除了中间变量 $X_{t-1},\cdots,X_{t-k+1}$ 带来的间接相关后的直接相关性，它是在已知序列值 $X_{t-1},\cdots,X_{t-k+1}$ 的条件下，X_t 与 X_{t-k} 间关系的度量。

定义：对于平稳 AR（p）序列，所谓滞后 k 偏自相关系数就是指在给定中间 $k-1$ 个随机变量的条件下，或者说，在剔除了中间 $k-1$ 个随机变量 $X_{t-1},\cdots,X_{t-k+1}$ 的干扰之后，X_{t-k} 对 X_t 影响的相关度量。用数学语言描述就是：

$$\rho_{X_t,X_{t-k}|X_{t-1},\cdots,X_{t-k+1}}=\frac{E\left[\left(X_t-\hat{E}X_t\right)\left(X_{t-k}-\hat{E}X_{t-k}\right)\right]}{E\left[\left(X_{t-k}-\hat{E}X_{t-k}\right)^2\right]}$$

常用 AR 模型偏自相关系数公式有如下两种。

（1）AR（1）模型

$$X_t=a_1X_{t-1}+\varepsilon_t$$

$$\phi_{11} = \rho_1 = a_1,$$

$$\phi_{kk} = 0, (k \geqslant 2)$$

（2）AR（2）模型

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \varepsilon_t$$

$$\phi_{11} = \rho_1 = \frac{a_1}{1 - a_2},$$

$$\phi_{22} = \frac{\rho_2 - (\rho_1)^2}{1 - (\rho_1)^2} = a_2, \ (\rho_2 = a_1\rho_1 + a_2\rho_0)$$

$$\phi_{kk} = 0, (k \geqslant 3)$$

3.4 AR 模型的 MATLAB 实现

在 MATLAB 中，自相关函数和偏自相关函数可以利用自带函数方便地画出。

3.4.1 自相关函数图

绘制自相关函数图，可以用 autocorr 函数来显示，具体用法为：

autocorr(y)

autocorr(y,numLags)

autocorr(y,numLags,numMA,numSTD)

acf = autocorr(y)

acf = autocorr(y,numLags)

acf = autocorr(y,numLags,numMA,numSTD)

[acf,lags,bounds] = autocorr(___)

（1）输入变量含义

- y——观测时间序列变量，最后一个元素是最新的观察值，指定缺失值用 NaN 代替。
- NumLags——样本的滞后数，默认值是 min（[20，T−1]），其中 T 是有效样本大小。
- NumMA——理论 MA 模型的滞后数。
- NumSTD——置信区间标准误倍数。

（2）输出变量含义

- acf——样本的自相关数，长度等于 NumLags+1，滞后阶数从 0 开始，0 阶时自相关系数为 1。
- lags——估计 ACF 的滞后阶数。
- bounds——估计的置信区间上下界。

例如，我们生成一组符合 MA（2）模型关于 MA 模型的具体内容在后面章节会介绍，维数为 1000 的数据，计算其偏自相关数及置信区间上下界可得：

```
>> [acf,lags,bounds] = autocorr(y,20,2);
>> bounds

bounds =

    0.0757
   -0.0757
```

可知其置信区间为（−0.0757,0.7573）

绘制其自相关图：

```
>>autocorr(y)
```

可得到如图 3-2 所示的自相关图，通过该图可以观察到自相关图像二阶截尾，符合 MA（2）模型。

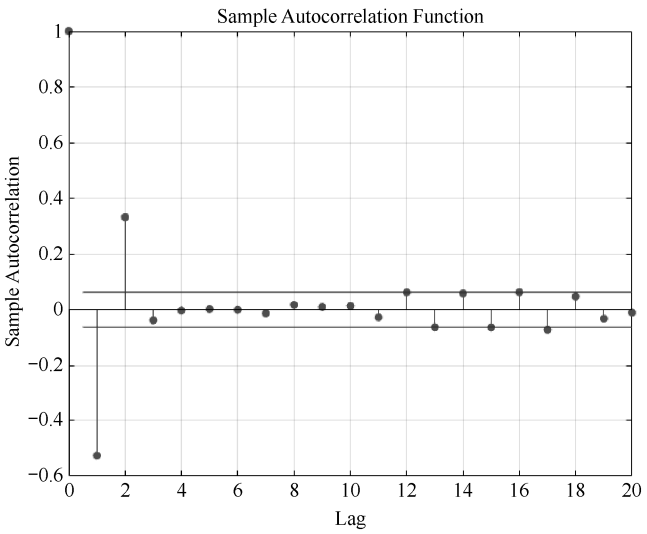


图 3-2　自相关图

3.4.2　样本偏自相关函数

样本的偏自相关数可用 parcorr 函数来计算，具体用法为：

parcorr(y)

parcorr(y,numLags,numAR,numSTD)

pacf = parcorr(y)

pacf = parcorr(y,numLags)

pacf = parcorr(y,numLags,numAR,numSTD)

[pacf,lags,bounds] = parcorr(___)

（1）输入变量含义

- y——观测时间序列变量，最后一个元素是最新的观察值，指定缺失值用 NaN 代替。
- NumLags——样本的滞后数，默认值是 min（[20，T−1]），其中 T 是有效样本大小。
- NumAR——理论 AR 模型的滞后数。

- NumSTD——置信区间标准误倍数。

（2）输出变量含义

- pacf——样本的偏自相关数，长度等于 NumLags+1，滞后阶数从 0 开始，0 阶时自相关系数为 1。
- lags——估计 PACF 的滞后阶数。
- bounds——估计的置信区间上下界。

例如，生成一组符合 AR（2）模型，维数为 1000 的数据，并计算其偏自相关数及置信区间上下界，可调用该函数直接实现：

```
>> [partialACF,lags,bounds] = parcorr(y,20,2);
>> bounds

bounds =

    0.0633
   -0.0633
```

可知其置信区间为（−0.0633,0.0633），绘制其偏自相关数图为：

```
>> parcorr(y)
```

可得到如图 3-3 所示的偏自相关图像二阶截尾，符合 AR（2）模型。

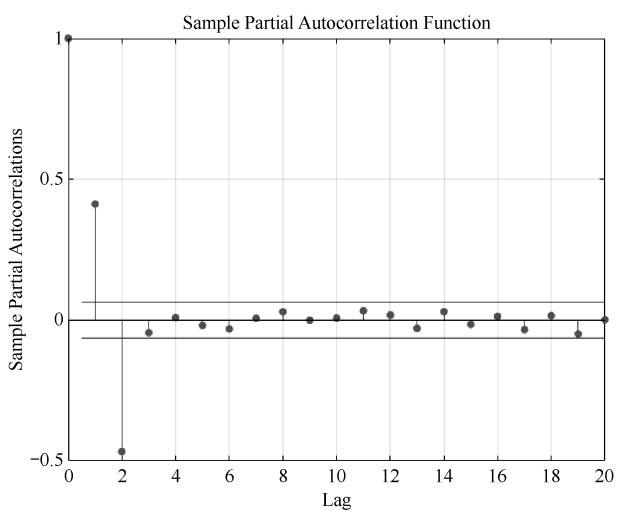


图 3-3　偏自相关图像

3.4.3 AR 模型函数

AR 模型的实现可以用函数 ar 来实现，具体用法为：

m = ar(y,n)

[m,ref1] = ar(y,n,approach,window)

m= ar(y,n,Name,Value)

m= ar(y,n,___,opt)

（1）输入变量含义

- y——需要输入的时间序列数据。
- n——估计的 AR 模型的阶数。
- approach——用来估计 AR 模型系数的方法，有如下选择：
  - 'fb'——默认选项，采用向前−向后逼近法。
  - 'burg'——基于网格的 Burg 方法。
  - 'gl'——几何网格逼近。
  - 'ls'——最小二乘逼近。
  - 'yw'——Yule-Walker 逼近。
- window——测量时间段以外的数据信息，有如下选择：
  - 'now'——该选项是在除了'yw'方法外其他方法的默认选项，只使用选用时间段内的数据去构建回归向量。
  - 'pow'——后窗口。最后的数值用 0 替代。
  - 'ppw'——前-后窗口。在 Yule-Walker 逼近中使用。
  - 'prw'——前窗口。丢失的前值用 0 替代。
- opt——可以设定特殊估计选项，有如下选择：

- data offsets。
- covariance handling。
- estimation approach。
- estimation window。

- Name——指定可选的由逗号分隔的参数对。Name 是参数名称，Value 是相应的值，名称必须出现在单引号' '内。可以按任意顺序指定几个名称和值对参数，如 Name1，Value1，…，NameN，ValueN。model = ar(y, 4, 'ls', 'Ts', Ts, 'IntegrateNoise', true)。
  - 'Ts'——指定采样时间的正标量。当使用的是 double 变量而不是 IDDATA 对象时使用。
  - 'IntegrateNoise'——布尔值，指定噪声源是否包含积分器。用于创造“ARI”结构模型 $Ay = \frac{e}{(1-z^{-1})}$

（2）输出变量含义

- m——输出的时间序列模型。
- refl——一个 2×2 数组。第一行存储反应系数，第二行存储相应的损失函数值。refl 的第一列是零阶模型，refl 的（2,1）元素是时间序列本身的范数。

3.5 AR 模型的应用实例

利用 AR 模型可以对股价的走势进行建模，从而便于研判股价未来的走势。比如对某股票从 2016 年 12 月～2017 年 5 月的收盘价进行 AR 模型的数值试验，画其自相关函数图，可以用如下脚本实现。

（1）读取股票数据

```
clc, clear all, close all
[closenum,date]=xlsread('data','Sheet1','A2:B101');
```

```
n = length(closenum);
a = zeros(n,1);
%转换日期数据格式
for i = 1:n
   a(i) = datenum(cell2mat(date(i)));
end
a = datestr(a);
a = datevec(a);
time = datetime(a);
```

（2）画自相关图像及偏自相关函数图

```
Figure
autocorr(closenum);
figure
parcorr(closenum);
```

某股价序列自相关函数图如图 3-4 所示。

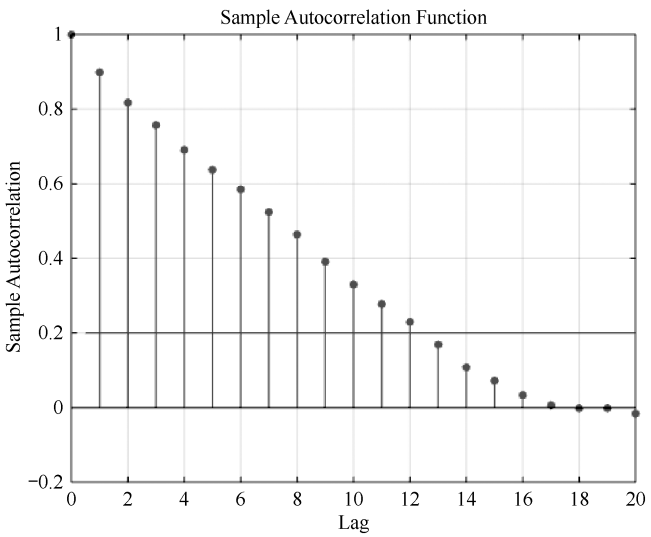


图 3-4　某股价序列的自相关函数图

通过图 3-4 可以看出该时间序列的自相关函数图是拖尾的。再利用 MATLAB 函数画出其偏自相关函数图像，如图 3-5 所示，可以看到偏自相关函数图像是 1 阶截尾的，所以使用 AR（1）模型进行拟合。

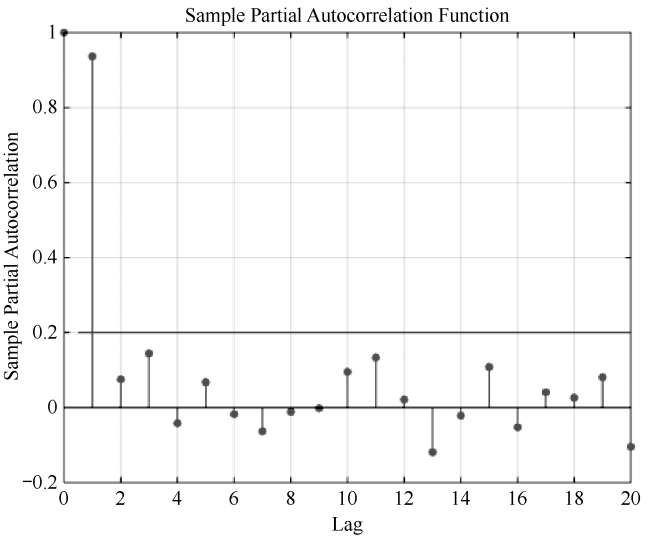


图 3-5　某股价序列的偏自相关函数图像

（3）建模

```
ar(closenum,1)
```

这样就可以得到该股的价格的 AR 模型：

```
Discrete-time AR model:  A(z)y(t) = e(t)
A(z) = 1 - z^-1
```

3.6　小结

本章介绍了平稳时间序列的自回归模型——AR 模型。该模型是用自身做回归变量的过程，即用前期若干时刻的随机变量的线性组合来描述以后某时刻随机变量的线性回归模型。我们介绍了该模型的原理、平稳性判别方法及基础统计性质。此外非常重要的是，我们给出了自相关函数及偏自相关函数的定义及其在 MATLAB 中的应用方法，在 MATLAB 中利用函数 autocorr 及 parcorr，可以方便地画出样本的自相关函数图及偏自相关函数图，方便时间序列建模的分析。在 MATLAB 中可以直接利用函数 ar 建立 AR 模型，我们在 3.5 节给出了 AR 模型在股票市场中的一个建模应用。

参考文献

[1] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[2] Hamilton. James，夏晓华. 时间序列分析. 北京：中国人民大学出版社，2015.

[3] 王黎明，王连，杨楠. 应用时间序列分析. 上海：复旦大学出版社，2009.

[4] 孙祝岭. 时间序列与多元统计分析. 上海：上海交通大学出版社，2016.

[5] 胡永宏，王振龙. 应用时间序列分析. 北京：科学出版社，2007.

4 滑动平均模型——MA 模型

本章讨论一类在金融收益率建模中很有用的简单模型——滑动平均模型（Moving-Average，MA），MA 可以看成参数受某种限制的无穷阶 AR 模型。

4.1 MA 模型的定义

4.1.1 由 AR 模型推导 MA 模型

我们考虑一个无穷阶的 AR 模型：

$$X_t = a_0 + a_1 X_{t-1} + \cdots + \varepsilon_t$$

由于有无穷个参数，在实际中这样的 AR 模型没有意义，想使这样的模型有实际用途，需要假定系数 a_i 满足某种限制，使得它们可由有限个参数决定，这种想法的一个特殊情形为：

$$X_t = a_0 - b_1 X_{t-1} - \cdots + \varepsilon_t$$

其中系数只依赖于单个参数 b_1，$a_i = -b_1^i$，$i \geqslant 1$。要使模型是平稳的，b_1 必须是绝对值小于 1 的，否则 b_1^i 序列本身将发散。因为 $|b_1| < 1$，故当 $i \to \infty$ 时，有 $b_1^i \to 0$。从而 X_{t-i} 对 X_t 的影响随 i 的增加以指数速度衰减。这一点是合理的，因为如果平稳序列 X_t 对它的延迟值 X_{t-i} 有依赖的话，这种依赖程度应随时间衰减。

我们将模型变化一下：

$$X_t + a_1 X_{t-1} + \cdots = a_0 + \varepsilon_t$$

对 X_{t-1} 的模型为：

$$X_{t-1} + b_1 X_{t-2} + b_1^2 X_{t-3} + \cdots = a_0 + \varepsilon_{t-1}$$

两边乘以b_1，然后两式相减，得到：

$$X_t = a_0(1-b_1) + \varepsilon_t - b_1\varepsilon_{t-1}$$

此式说明除去常数项外，X_t是两个扰动ε_t和ε_{t-1}的加权平均。因此，此模型称为1阶MA模型，简称为MA（1）模型。

4.1.2 MA模型的定义

MA（1）模型的一般形式为：

$$X_t = c_0 + \varepsilon_t - b_1\varepsilon_{t-1} \text{ 或 } r_t = c_0 + (1-b_1L)\varepsilon_t$$

其中c_0是一个常数；$\{\varepsilon_t\}$是一个白噪声序列。类似地，MA（2）模型的形式为：

$$X_t = c_0 + \varepsilon_t - b_1\varepsilon_{t-1} - b_2\varepsilon_{t-2}$$

MA（q）模型为：

$$X_t = c_0 + \varepsilon_t - b_1\varepsilon_{t-1} - b_2\varepsilon_{t-2} - \cdots b_q\varepsilon_{t-q}$$

或

$$X_t = c_0 + (1-b_1L - \cdots - b_qL^q)\varepsilon_t \text{，其中 } q>0$$

记$\beta(u) = c_0 + 1 - b_1u - \cdots - b_qu^q$，则MA（$q$）模型可用$X_t = c_0 + \beta(L)\varepsilon_t$表示。

4.2 MA模型的性质

4.2.1 平稳性

MA模型总是弱平稳的，因为它们是白噪声序列的有限线性组合，其前两阶矩是不随时间变化的。例如MA（1）模型，对这个模型两端取期望可得$E(X_t) = c_0$，这不随时间变化。两端取方差，我们有：

$$\text{Var}(X_t) = \sigma_\varepsilon^2 + b_1^2\sigma_\varepsilon^2 = (1+b_1^2)\sigma_\varepsilon^2$$

这里我们用到ε_t与ε_{t-1}的不相关性，$\text{Var}(X_t)$也不随时间变化。这些讨论对一般

的 MA（q）模型也适用，因此我们得到两个一般性质：第一，MA 模型的常数项就是序列的均值（也即 $E(X_t)=c_0$）；第二，MA（q）模型的方差为：

$$\mathrm{Var}(X_t)=(1+b_1^2+b_2^2+\cdots+b_q^2)\sigma_\varepsilon^2$$

4.2.2 自相关函数

为简单起见，假定 MA（1）模型中 $c_0=0$。对两端乘以 X_{t-l}，我们有：

$$X_{t-l}X_t=X_{t-l}\varepsilon_t-b_1X_{t-l}\varepsilon_{t-1}$$

取期望，得到 $\gamma_1=-b_1\sigma_\varepsilon^2$，且 $l>1$ 时，$\gamma_l=0$

利用上述结果，并注意到 $\mathrm{Var}(X_t)=(1+b_1^2)\sigma_\varepsilon^2$，我们有：

$$\rho_0=1,\rho_1=\frac{-b_1}{1+b_1^2},\rho_l=0\text{，其中 }l>1$$

因此，对 MA（1）模型，间隔为 1 的 ACF 不为 0，但所有间隔大于 1 的 ACF 都是 0。换言之，MA（2）模型的 ACF 在间隔为 1 以后是截尾的。对于 MA（2）模型，自相关系数是：

$$\rho_1=\frac{-b_1+b_1b_2}{1+b_1^2+b_2^2},\rho_2=\frac{-b_2}{1+b_1^2+b_2^2},\rho_l=0\text{，其中 }l>2$$

这时，在间隔为 2 以后截尾，这个性质可推广到其他 MA 模型。对 MA（q）模型，其 ACF 在间隔为 q 时不为 0，但对 $l>q$，$\rho_l=0$。因此，MA（q）序列只与其前 q 个延迟值线性相关，从而它是一个“有限记忆”的模型。

4.2.3 可逆性

此外不同的 MA 模型可能具有完全相同的自相关系数和偏自相关系数，为了利用自相关系数和偏自相关系数来识别 MA 模型，要求给定一个自相关函数能够对应唯一的 MA 模型，这就要求我们给模型增加约束条件，这个约束条件称为 MA 模型的可逆性条件。

可逆性定义：若一个 MA 模型能够表示成为收敛的 AR 模型形式，那么该 MA 模

型称为可逆 MA 模型，即可以保证一个自相关系数列唯一对应一个可逆 MA 模型。

将零均值 MA（1）模型改写为 $\varepsilon_t = X_t + b_1\varepsilon_{t-1}$，重复迭代可以得到：

$$\varepsilon_t = X_t + b_1 X_{t-1} + b_1^2 X_{t-2} + b_1^3 X_{t-3} + \cdots$$

该等式表明当前的扰动 ε_t 是现在和过去收益率序列的线性组合。从直观上看，随着 j 的增加，b_1^j 应该趋于零，因为遥远的过去收益率对当前的扰动应该几乎没有影响。因此，要使 MA（1）模型看起来是合理的，我们应该要求 $|b_1|<1$。这样的 MA（1）模型称为可逆的。如果 $|b_1|=1$，则 MA（1）模型是不可逆的。

进一步更严谨的定义：若 $\beta(u)=0$ 的根均在单位圆外，称 MA（q）模型为可逆的 MA（q）模型，$\{X_t\}$ 为可逆的 MA（q）序列。

定义：称 $\{b \mid \beta(u)=0$ 的根在单位圆外，$b^T=(b_1,b_2,\cdots,b_q)\in R^p\}$ 为 MA（q）模型的可逆域。

4.2.4 MA（1）模型可逆性分析

我们讨论 MA（1）模型：$x_t = \varepsilon_t - \theta\varepsilon_{t-1}$

（1）模型的可逆域 $=\{\theta \mid 0<|\theta|<1\}$；

（2）逆转形式即用 $\{X_t\}$ 的组合表示 ε_t，或者说从模型中解出 ε_t。具体的模型可逆性分析如图 4-1 所示。

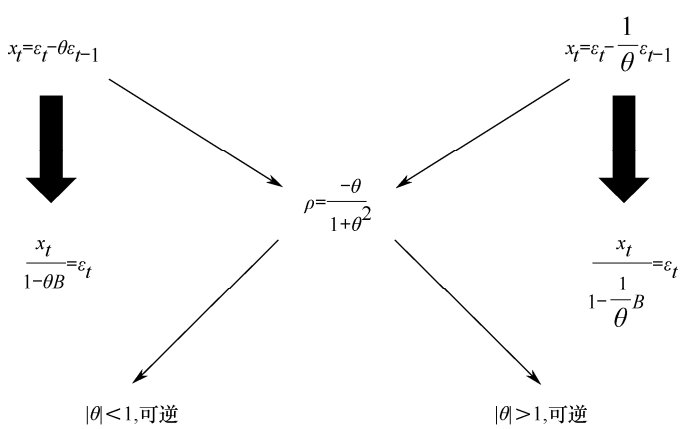


图 4-1　MA 模型可逆性分析

$$\varepsilon_t = \frac{1}{1-\theta L}\varepsilon_t = \sum_{j=0}^{+\infty}(\theta L)^j X_t = \sum_{j=0}^{+\infty}\theta^j X_{t-j}\ ;$$

如图 4-1 所示，MA（q）模型可逆的充要条件是：MA（q）模型的特征根都在单位圆内 $|u_i|<1,\forall i$，即算子多项式的根都在单位圆外 $\left|\frac{1}{u_i}\right|>1,\forall i$。

4.3 MA 模型的应用实例

在 MATLAB 中建立 MA 模型，与我们后面会介绍的 ARIMA 模型使用的是一个函数，因此我们在后文中再详细解释该函数，此处仅给出该方法的一个数值实例，对 MA 模型建模，可以用如下脚本实现。

（1）创建建模数据

```
Mdl = arima('MA',{0.5,-0.3},'Constant',0,'Variance',0.01);
rng(5);
Y = simulate(Mdl,130);
figure(1)
plot(Y,'LineWidth',1);
title('MA 数值实验数据图像')
```

创建的数据图像如图 4-2 所示。

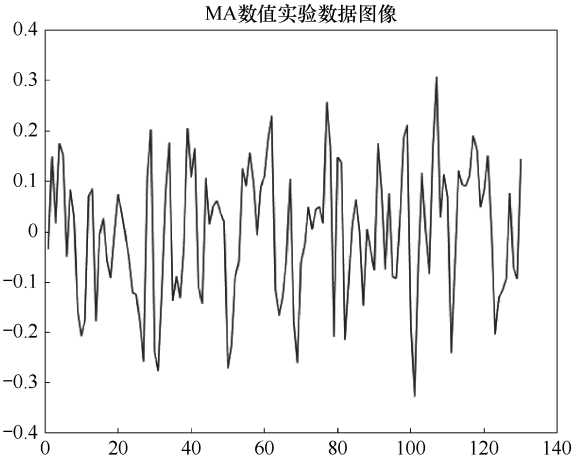


图 4-2　MA 数值实验数据图像

（2）画自相关及偏自相关函数图

```
figure(2)
subplot(2,1,1)
autocorr(Y);
title('数据的自相关图像')
subplot(2,1,2)
parcorr(Y);
title('数据的偏自相关图像')
```

如图 4-3 所示是数据的自相关及偏自相关函数图。

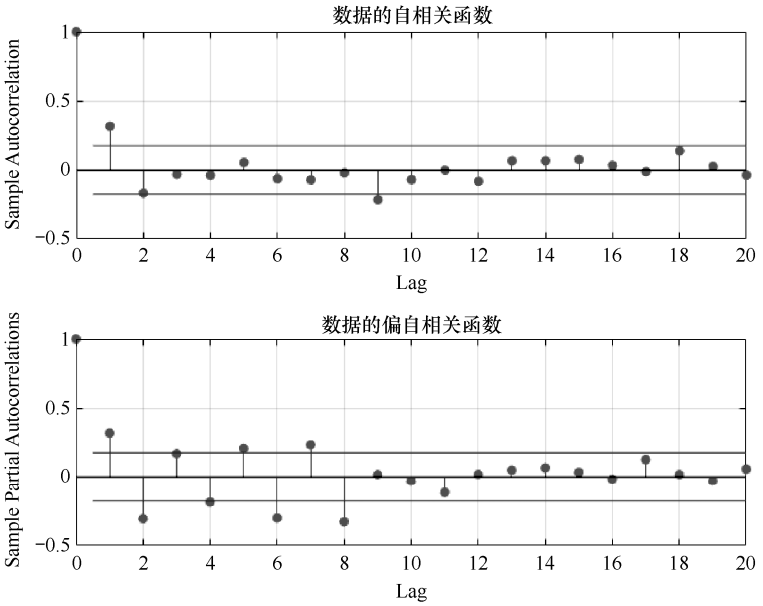


图 4-3　MA 数据的自相关及偏自相关函数图

（3）建模

```
ToEstMdl = arima('MALags',1:2);
fit = estimate(ToEstMdl, Y);
```

这样可以得到该数据的 MA 模型：

```
fit = estimate(ToEstMdl, Y);
```

```
ARIMA(0,0,2) Model:
--------------------
Conditional Probability Distribution: Gaussian

                          Standard          t
 Parameter      Value         Error      Statistic
-----------  -----------  ------------  -----------
 Constant    -0.00336305    0.0137197     -0.245125
   MA{1}       0.626133    0.0912921       6.85857
   MA{2}      -0.235414    0.0879128      -2.67781
 Variance      0.0121946   0.00173957       7.01008
```

4.4 小结

本章介绍了滑动平均模型——MA 模型。该模型可以看成是参数受某种限制的无穷阶 AR 模型，是由不同阶滞后的白噪声来拟合某一时刻的序列数据。我们介绍了其基本性质，包括平稳性、自相关函数及可逆性。可逆性是说如果一个 MA 模型能够表示成收敛的 AR 模型形式，它就是可逆的，该条件可以保证一个自相关系数列对应唯一一个可逆的 MA 模型。我们在 4.3 节给出了该方法的应用，由于在 MATLAB 中使用的是 arima 函数，具体函数的介绍我们将在后续的章节中详细说明。

参考文献

[1] Jonathan D.Cryer, Kung-Sik Chan. Time Series Analysis With Applications in R (Second Edition): New York, Springer, 2008.

[2] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[3] 王黎明，王连，杨楠. 应用时间序列分析. 上海：复旦大学出版社，2009.

[4] 胡永宏，王振龙. 应用时间序列分析. 北京：科学出版社，2007.

[5] 黄红梅. 应用时间序列分析. 北京：清华大学出版社，2016.

[6] 孙祝岭. 时间序列与多元统计分析. 上海：上海交通大学出版社，2016.

5 自回归滑动平均模型——ARMA 模型

在有些应用中，我们需要高阶的 AR 或 MA 模型才能充分地描述数据的动态结构。这样就有很多参数要估计，问题变得繁琐了。为了克服这个困难，人们提出了自回归滑动平均（ARMA）模型。基本思想是把 AR 和 MA 模型的想法结合在一个紧凑的形式中，使得模型保持较少的参数。

5.1 ARMA 模型介绍

5.1.1 ARMA 模型的定义

定义：假设时间序列 $\{X_t\}$ 适合：

$$X_t - a_0 - a_1 X_{t-1} - \cdots - a_p X_{t-p} = \varepsilon_t - b_1 \varepsilon_{t-1} - \cdots - b_q \varepsilon_{t-q}$$

$\{\varepsilon_t\}$ 为白噪声序列，$\forall s < t$，有 $EX_s \varepsilon_t = 0$，称此模型为自回归滑动平均模型，记为 ARMA（p,q）模型；称适合此模型的 $\{X_t\}$ 为 ARMA（p,q）序列。若 $a_0 = 0$，则称之为中心化的 ARMA（p,q）模型。

注意：模型可化为 $\alpha(L)X_t = \beta(L)\varepsilon_t$，一般需假定 $\alpha(u)$、$\beta(u)$ 无公共根。

$\alpha(L) = 1 - a_1 L - \cdots - a_p L^p$，为 p 阶自回归系数多项式。

$\beta(B) = 1 - b_1 L - \cdots - b_q L^q$，为 q 阶移动平均系数多项式。

显然当 $q = 0$ 时，ARMA（p,q）模型就退化成了 AR（p）模型；

当 $p = 0$ 时，ARMA（p,q）模型就退化成了 MA（q）模型。

所以，AR（p）模型和 MA（q）模型实际上是 ARMA（p,q）模型的特例，它们都统称为 ARMA（p,q）模型。而 ARMA（p,q）模型的统计性质也正是 AR（p）模型和 MA（q）模型统计性质的有机组合。

5.1.2 ARMA 模型的平稳条件

类似于 AR（p）模型平稳性的分析，容易推导出 ARMA（p,q）模型的平稳条件是：$\alpha(L)=0$ 的根都在单位圆外。也就是说，ARMA（p,q）模型的平稳性完全由其自回归部分的平稳性所决定。

同理，可以推导出 ARMA（p,q）模型的可逆条件和 MA（q）模型的可逆条件完全相同：当 $\beta(L)=0$ 的根都在单位圆外时，ARMA（p,q）模型可逆。

当 $\alpha(L)=0,\beta(L)=0$ 的根都在单位圆外时，ARMA（p,q）模型称为平稳可逆模型，这是一个由它的自相关系数唯一可以识别的模型。

5.2 ARMA 模型的性质

5.2.1 均值

对于一个非中心化平稳可逆的 ARMA（p,q）模型：

$$X_t=a_0+a_1X_{t-1}+\cdots+a_pX_{t-p}+\varepsilon_t-b_1\varepsilon_{t-1}-\cdots-b_q\varepsilon_{t-q}$$

两边同求均值，有：

$$EX_t=\frac{a_0}{1-a_1-\cdots-a_p}$$

5.2.2 自协方差函数

ARMA（p,q）模型 $X_t-a_1X_{t-1}-\cdots-a_pX_{t-p}=\varepsilon_t-b_1\varepsilon_{t-1}-\cdots-b_q\varepsilon_{t-q}$ 的序列 $\{X_t\}$ 的自协方差函数为：

$$\sum_{j=0}^{p}\sum_{i=0}^{p}a_j a_i \gamma_{k-i+j}=\begin{cases}\sigma^2(1+\sum_{j=1}^{q}b_j^2), & k=0\\ \sigma^2(1+\sum_{j=1}^{q-k}b_j b_{j+k}), & 1\leqslant k\leqslant q\\ 0, & k>q\end{cases}$$

令 $X_t-a_1X_{t-1}-\cdots-a_pX_{t-p}=Y_t$，则 $\{Y_t\}$ 是一个 MA（q）序列，故：

$$\gamma_k(Y)=\begin{cases}\sigma^2(1+\sum_{j=1}^{q}b_j^2), & k=0\\ \sigma^2(-b_k+\sum_{j=1}^{q-k}b_j b_{j+k}), & 1\leqslant k\leqslant q\\ 0, & k>q\end{cases}$$

另外 $\gamma_k(Y)=EY_tY_{t+k}=\sum_{j=0}^{p}\sum_{i=0}^{p}a_j a_i EX_{t-j}X_{t+k-i}=\sum_{j=0}^{p}\sum_{i=0}^{p}a_j a_i\gamma_{k-i+j}$ 所以成立。

计算 ARMA（p,q）序列的协方差函数有递推公式如下。

由 $X_t-a_1X_{t-1}-\cdots-a_pX_{t-p}=\varepsilon_t-b_1\varepsilon_{t-1}-\cdots-b_q\varepsilon_{t-q}$ 两边乘以 X_{t-k}，得：

$$X_tX_{t-k}=a_1X_{t-1}X_{t-k}+a_2X_{t-2}X_{t-k}+\cdots+\varepsilon_tX_{t-k}-b_1\varepsilon_{t-1}X_{t-k}-\cdots-b_q\varepsilon_{t-q}X_{t-k}$$

两边取期望，得：

$$EX_tX_{t-k}=a_1EX_{t-1}X_{t-k}+a_2EX_{t-2}X_{t-k}+\cdots+a_pEX_{t-p}X_{t-k},k>q$$

所以 $\gamma_k=a_1\gamma_{k-1}+a_2\gamma_{k-2}+\cdots+a_p\gamma_{k-p},k>q$，即 $\alpha(L)\gamma_k=0,k>q$

5.2.3 ARMA（1,1）模型的性质

我们考虑 ARMA（1,1）模型：

$$X_t=aX_{t-1}+\varepsilon_t+b\varepsilon_{t-1}$$

（1）平稳域 $=\{a|0<|a|<1\}$；可逆域 $=\{b|0<|b|<1\}$

（2）自协方差函数

$k=0$ 时，$\gamma_0-2a\gamma_1+a^2\gamma_0=\sigma^2(1+b^2)$

$k=1$ 时，$-a\gamma_0-a^2\gamma_1+a^2\gamma_1+\gamma_1=b\sigma^2$

$k>1$ 时，$\gamma_k=a\gamma_{k-1}$

解得：$\gamma_0=\dfrac{1+2ab+b^2}{1-a^2}\sigma^2;\gamma_1=\dfrac{(1+ab)(a+b)}{1-a^2}\sigma^2$

自相关函数 $\rho_0=1,\rho_1=\dfrac{(1+ab)(a+b)}{1+2ab+b^2},\rho_k=a\rho_{k-1},k\geqslant 2$

（3）平稳解（也称传递形式）

对原模型移项，得：

$$(1-aL)X_t=b\varepsilon_{t-1}+\varepsilon_t$$

形式上可化为：

$$\begin{aligned}X_t&=\frac{b\varepsilon_{t-1}+\varepsilon_t}{1-aL}=\sum_{j=0}^{+\infty}(aL)^j(b\varepsilon_{t-1}+\varepsilon_t)=b\sum_{j=0}^{+\infty}a^j\varepsilon_{t-j-1}+\sum_{j=0}^{+\infty}a^j\varepsilon_{t-j}\\&=\varepsilon_t+b\sum_{j=0}^{+\infty}a^j\varepsilon_{t-j-1}+\sum_{j=1}^{+\infty}a^j\varepsilon_{t-j}=\varepsilon_t+b\sum_{i=1}^{+\infty}a^{i-1}\varepsilon_{t-i}+\sum_{j=1}^{+\infty}a^j\varepsilon_{t-j}\\&=\varepsilon_t+\sum_{i=1}^{+\infty}(a+b)a^{i-1}\varepsilon_{t-i}\end{aligned}$$

验证是解略。易证 X_t 是平稳解。

（4）逆转形式：

$$\varepsilon_t=\frac{1}{1+bL}(X_t-aX_{t-1})=X_t+\sum_{i=1}^{+\infty}(-1)^i(a+b)b^{i-1}X_{t-i}$$

5.3 ARMA 模型的图像定阶

观察自相关图与偏相关图可以确定序列的 ARMA（p，q）模型的具体形式。首

先，需要明确这样 2 对概念：

第一，自回归过程与移动平均过程。自回归由序列的滞后变量的线性组合以及白声噪（符合 0 均值固定方差的随机干扰项）相加而成；移动平均过程由白声噪的线性组合构成。

第二，拖尾和截尾。这一对概念前者指 ACF 或者 PACF 呈几何衰减（指数式衰减或者正弦式衰减）；后者指 ACF 或者 PACF 在某一阶之前明显不为 0，之后突然接近或者等于 0。其实，从字面上也很好理解，拖尾就是拖拖拉拉，截尾就是抽刀断水。

其次是对 ARMA 模型的分解。

（1）AR（p）模型

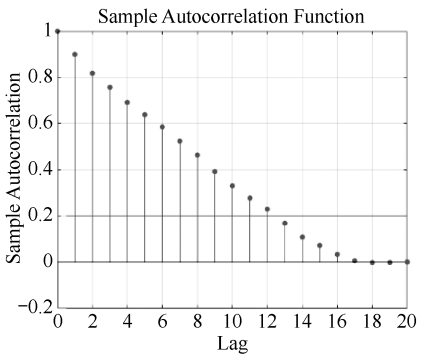
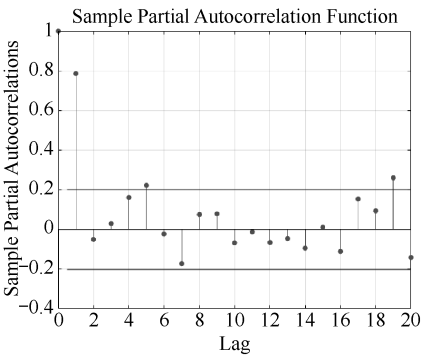
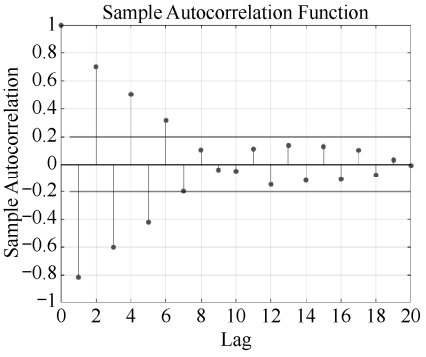
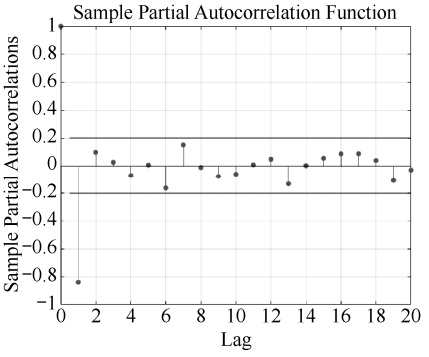
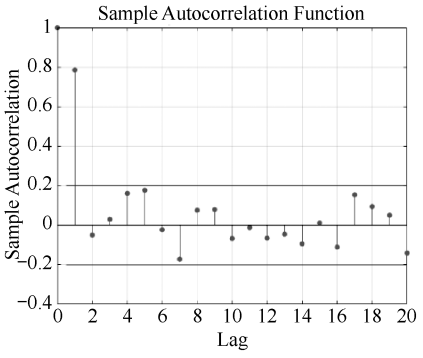
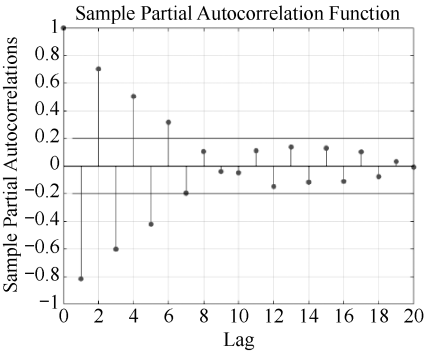
从自相关函数 ACF 来看，在自回归方程的基础上可以很简单地构造自相关系数，对于平稳时间序列（注意这一前提条件，如果放开这一条件图形将会很难识别），α 为自回归系数，且$|\alpha|<1$，所以当$\alpha>0$ 时，ACF 呈现为指数式衰减至 0。当$\alpha<0$ 时，ACF 则正负交替呈指数衰减至 0，整体表现则是正弦式衰减；从偏相关函数 PACF 来看，这就相当明显了，因为 PACF 与自回归方程的形式完全一样，只是自回归方程只有滞后 p 期，而 PACF 则有更多的滞后项。于是，很明显，当 $k\leqslant p$ 时,偏相关系数不等于 0；当 $k>p$ 时，偏相关系数等于 0，明显呈现出截尾现象。

（2）MA（q）模型

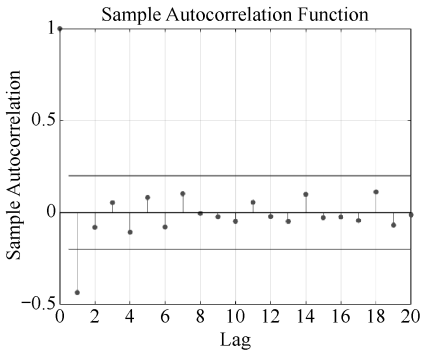
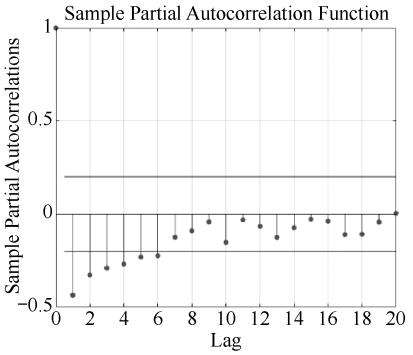
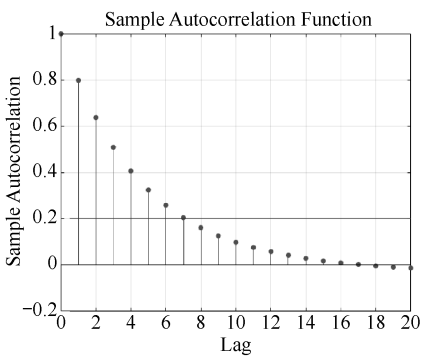
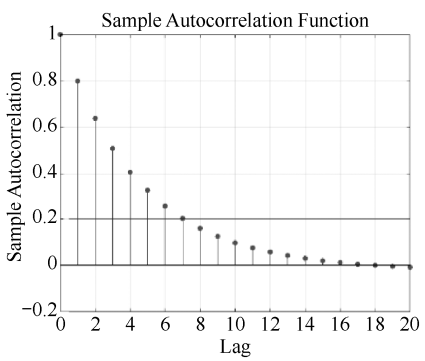
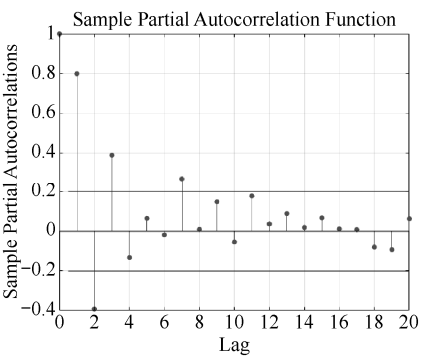
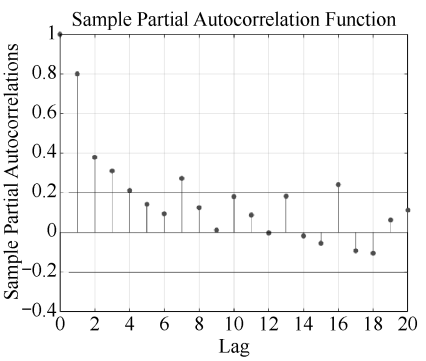
从自相关函数 ACF 来看，在移动平均方程的基础上也可以很简单地构造自相关函数，这时候的自相关函数为分段函数，当 $k\leqslant p$ 时偏相关系数不等于 0，当 $k>q$ 时，偏相关系数等于 0，明显呈现出截尾现象；从偏相关函数 PACF 来看，任何一个可逆的 MA（q）过程都可以转换成一个无限阶、系数按几何衰减的 AR 过程（将白噪声替换为序列的滞后形式即可），呈现拖尾现象。与 AR（p）不同的是，当 $q>0$（q 为移动平均系数）时，PACF 呈现为交替式正弦衰减。当 $q<0$ 时，PACF 则呈指数衰减至 0。ARMA（p，q）模型则是两者的结合，实际判别 p、q 值时还是比较依赖经验的。

表 5-1　总结了常见的 ARMA 模型定阶图像。

表 5-1　模型的图像定阶

模　　型	自相关函数特征	偏自相关函数特征
AR（1） $X_t=a_1X_{t-1}+\varepsilon_t$	若 $a_1>0$，平滑地指数衰减 	若 $a_1>0$，k=1 时有正峰值然后截尾 
	若 $a_1<0$，正负交替地指数衰减 	若 $a_1<0$，k=1 时有负峰值然后截尾 
MA（1） $X_t=\varepsilon_t+b_1\varepsilon_{t-1}$	若 $b_1>0$，k=1 时有正峰值然后截尾 	若 $b_1>0$ 时，交替式指数衰减 

续表

模　　型	自相关函数特征	偏自相关函数特征
MA（1） $X_t=\varepsilon_t+b_1\varepsilon_{t-1}$	若 $b_1<0$，$k=1$ 时有负峰值然后截尾 	若 $b_1<0$ 时，负的平滑式指数衰减 
ARMA （1,1） $X_t=a_1X_{t-1}+$ $\varepsilon_t+b_1\varepsilon_{t-1}$	$k=1$ 有峰值然后按指数衰减  ($a_1>0$，$b_1>0$)  ($a_1>0$，$b_1<0$)	$k=1$ 有峰值然后按指数衰减  ($a_1>0$，$b_1>0$)  ($a_1>0$，$b_1<0$)

5.4 ARMA 模型的应用实例

ARMA 模型建模在 MATLAB 中使用的是 arima 函数，具体函数的使用方法我们将在下一章中进行讲述。利用 ARMA 模型可以对指数涨跌幅数据进行分析，比如使用上证 50 指数 2016 年 3 月 21 日—2018 年 1 月 17 日涨跌幅数据来建立 ARMA 模型，可以用如下脚本实现。

（1）读取指数数据

```
clc, clear all, close all
[pct,date]=xlsread('50pct_change','Sheet1','A2:B451');
N = length(pct);
a = zeros(N,1);
%处理日期数据格式
for i = 1:N
   a(i) = datenum(cell2mat(date(i)));
end
a = datestr(a);
a = datevec(a);
Dates = datetime(a);
```

（2）数据可视化

```
figure(1)
plot(Dates,pct);
title('上证 50 指数涨跌幅')
ylabel('涨跌幅')
```

如图 5-1 所示，可以看出该涨跌幅基本符合时间序列数据的要求，但是仍需要做平稳性数值检验。

（3）平稳性检验

我们采用 pp 检验来看该数据是否有单位根，如果不能拒绝原假设则说明指数序列存在单位根。

```
disp('使用 PP 检验，如果不能拒绝原假设则说明指数序列存在单位根')
```

```
[hp,hpValue,stat,cValue,reg] = pptest(pct,'model','TS')
```

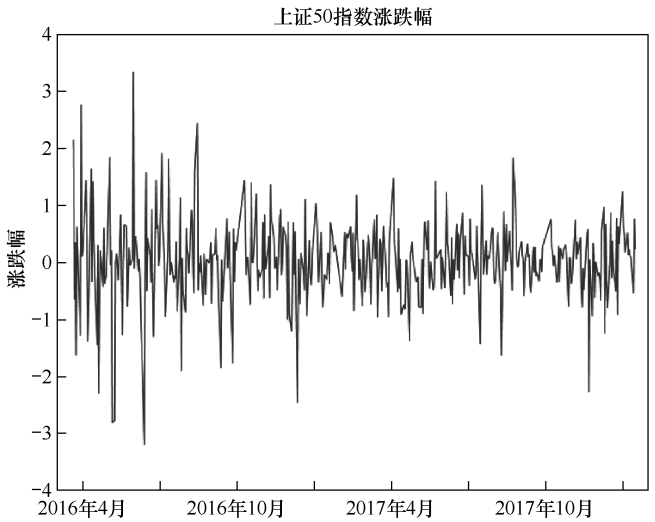


图 5-1　上证 50 指数涨跌幅

检验结果为：

```
hp =

  logical

   1


hpValue =

   1.0000e-03
```

该检验结果显示 hp=1，则说明该数据没有单位根是平稳数据，可以进行建模。

（4）绘制自相关及偏自相关函数图

```
figure(2)
subplot(2,1,1)
autocorr(pct);
title('指数涨跌幅的自相关图像')
```

```
subplot(2,1,2)
parcorr(pct);
title('指数涨跌幅的偏自相关图像')
```

图 5-2 展示了上证 50 指数涨跌自相关及偏自相关函数图像。

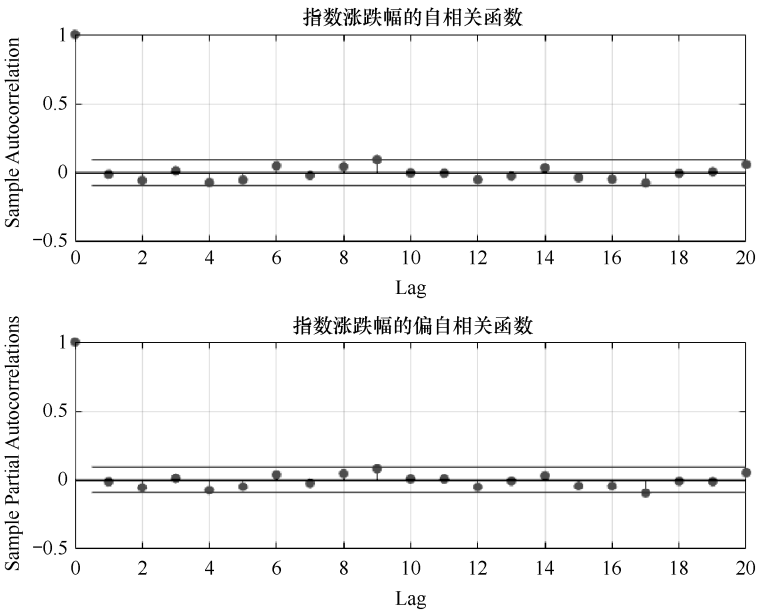


图 5-2　上证 50 指数涨跌自相关及偏自相关函数图像

（5）AIC 准则定阶

更精确的，我们使用 AIC 准则进行定阶，该函数的具体内容我们将在后面的章节讲述。定阶时因为各个定阶互不干扰，可以采用 MATLAB 自带的并行工具箱提高运算效率。

```
maxLags = 4;
AICSet = zeros(maxLags, maxLags);

parfor i = 1:maxLags% parfor 表示使用并行工具箱计算
for j = 1:maxLags
      mdl = arima('ARLags',[1:i],'MALags',[1:j]);
      [EstMdl, EstParamCov, logL, info] = estimate(mdl, pct,
      'display', 'off');
```

```
            AICSet(i, j) = aicbic(logL, length(info.X));

        end
        end

        % 画热度图来表示 AIC 数值的分布
        figure(3)
        heatmap(AICSet/1000);
        xlabel('MA Lags')
        ylabel('AR Lags')
        title('Akaike information criteria')

        [OptimalARLags, OptimalMALags] = find(AICSet==min(min(AICSet)));
        title(['Optimal AR and MA Lags are (' num2str(OptimalARLags) ','
num2- str(OptimalMALags) ')'])
```

图 5-3 是将 AIC 准则的分布绘制成热度图，竖轴表示 AR 的阶数，横轴表示 MA 的阶数，通过该图像可以直观地选择 AIC 准则最小的建模，建立 ARMA（3,3）模型。

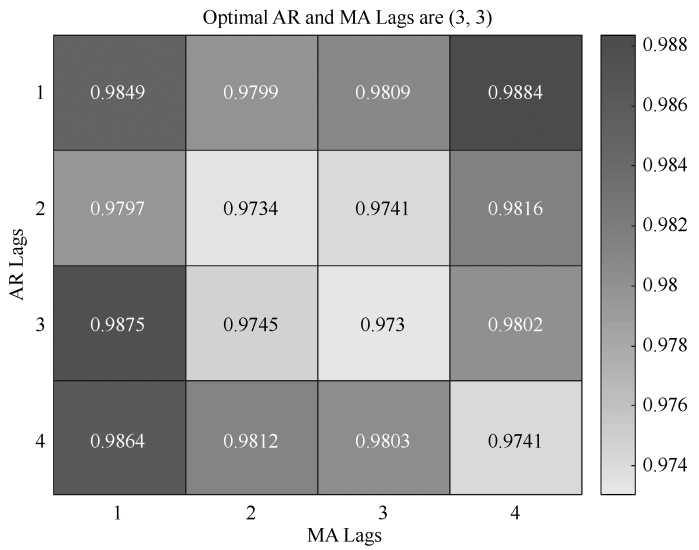


图 5-3　AIC 准则定阶分布图

（6）建立模型

```
        mdl = arima( OptimalARLags,0, OptimalMALags);
```

```
fit = estimate(mdl, pct);
```

建立模型可以得到如下结果：

```
ARIMA(3,0,3) Model:
  --------------------
  Conditional Probability Distribution: Gaussian

                              Standard         t
   Parameter       Value         Error      Statistic
  -----------   -----------   ------------   -----------
   Constant      0.016101     0.0242378      0.664295
      AR{1}     -0.862923      0.132241      -6.52538
      AR{2}      0.656269      0.231712       2.83226
      AR{3}      0.712036      0.114425       6.22275
      MA{1}      0.888515      0.116104       7.65274
      MA{2}     -0.694955      0.207444      -3.35008
      MA{3}     -0.792489      0.108523      -7.30252
   Variance      0.491066     0.0227658       21.5703
```

5.5 小结

本章我们介绍了自回归滑动平均模型——ARMA 模型。因为在某些应用中我们需要高阶的 AR 或 MA 模型才能充分地描述数据的动态结构。为了克服这个困难，人们提出了将 AR 模型和 MA 模型结合的 ARMA 模型。AR 模型和 MA 模型实际上是 ARMA 模型的特例。我们还介绍了 ARMA 模型的统计性质，该模型的统计性质是 AR 模型和 MA 模型统计性质的有机组合。在 5.3 节中我们给出了根据自相关图及偏自相关图对 ARMA 模型定阶的方法。看图主要是看图像是拖尾还是截尾，拖尾是指自相关图或偏自相关图呈几何衰减（指数式衰减或者正弦式衰减）；截尾指自相关图或偏自相关图在某一阶之前明显不为 0，之后突然接近或者等于 0。对于平稳时间序列数据，AR 模型的自相关函数呈现拖尾性，偏相关呈现截尾性。而 MA 模型正好相反，自相关函数呈现截尾性，偏相关呈现拖尾性。在 5.4 节中我们用上证 50 指数的涨跌幅数据给出了 ARMA 模型的应用实例。

参考文献

[1] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[2] 孙祝岭. 时间序列与多元统计分析. 上海：上海交通大学出版社，2016.

[3] 黄红梅. 应用时间序列分析. 北京：清华大学出版社，2016.

[4] Jonathan D.Cryer, Kung-Sik Chan. Time Series Analysis With Applications in R (Second Edition): New York, Springer, 2008.

[5] SerenaNg, PierrePerron(1995). Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag[J]. Publications of the American Statistical Association, 90(429):268-281.

6 非平稳序列的随机分析——ARIMA模型

ARIMA 模型全称为自回归积分滑动平均模型（Autoregressive Integrated Moving Average Model，ARIMA），是由博克思（Box）和詹金斯（Jenkins）于 20 世纪 70 年代初提出的著名时间序列预测方法，所以又称为 Box-Jenkins 模型。

6.1 ARIMA 模型的定义

ARIMA（p, d, q）称为自回归积分滑动平均模型，AR 是自回归，p 为自回归项；MA 为滑动平均，q 为滑动平均项数，d 为时间序列成为平稳时所做的差分次数。所谓 ARIMA 模型，是指将非平稳时间序列转化为平稳时间序列，然后仅对因变量的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型。ARIMA 模型根据原序列是否平稳以及回归中所含部分的不同，包括滑动平均过程（MA）、自回归过程（AR）、自回归滑动平均过程（ARMA）以及 ARIMA 过程。

ARIMA 模型是一类常用的随机时序模型，ARIMA（p, d, q）模型是指 d 阶差分后自相关最高阶数为 p、滑动平均最高阶数为 q 的模型，通常它包含 $p+q$ 个独立的未知系数。

如果该模型中有部分自相关系数或部分滑动平滑系数为零，即原 ARIMA（p, d, q）模型中有部分系数省缺了，那么该模型称为疏系数模型。

如果只是自相关部分有省缺系数，那么该疏系数模型可以简记为：

$$\text{ARIMA}((p_1,\cdots, p_m), d, q)$$

式中，$p_1,\cdots, p_m$ 为非零自相关系数的阶数。

如果只是滑动平滑部分有省缺系数，那么该疏系数模型可以简记为：

$$\text{ARIMA}(p,d,(q_1，\cdots，q_n))$$

式中，$q_1,\cdots，q_n$ 为非零移动平滑系数的阶数。

如果自相关和移动平滑部分都有省缺，可以简记为：

$$\text{ARIMA}((p_1,\cdots,p_m),d,(q_1,\cdots,q_n))$$

在实际操作中，疏系数模型时有应用。

6.2 ARIMA 模型的 MATLAB 实现

对于该模型，我们在 MATLAB 中使用 arima 和 estimate 两个函数来建立模型。

6.2.1 ARIMA 建模

创建时间序列模型采用 arima 函数，该函数能建立包括 MA 模型、AR 模型、ARMA 模型、ARIMA 模型及季节性时间序列模型。此外该函数还可以对已知系数或者使用数据估计出来的系数创建模型。默认情况下方差是正标量，但可以指定任何支持的条件方差模型。

Mdl = arima(p,D,q) 创建非季节性时间序列模型

Mdl = arima(Name,Value) 根据输入的条件创建指定的时间序列模型

输入变量含义如下。

- p——正整数，表示非季节性自回归阶数。
- D——非负整数，表示非季节性积分，即 ARIMA 模型中差分次数。
- q——正整数，表示非季节性移动平均指数。

可选择变量的含义如下。

- AR——默认值：元素为 NaN 的元胞向量。非季节性自回归模型多项式对应的

系数。当没有指定 ARLags 时，AR 是滞后 1，2，…阶数的多项式系数，当指定滞后阶数 ARLags 时，AR 是与滞后数等长的系数向量。

- ARLags——与 AR 系数相关的正整数滞后向量。默认值为与非季节性自回归多项式阶数相等的整数向量。
- Beta——在 ARIMA 条件均值模型中对应于回归分量的系数的实数向量。默认值：[ ]（没有对应于回归分量的回归系数）。
- Constant——线性时间序列中的常量系数。默认值：NaN。
- D——线性时间序列中非季节差分滞后算子多项式的程度。默认值：0（没有非季节性）。
- Distribution——过程的条件概率分布。默认值：高斯分布。
- MA——非季节性滑动平均模型的可以多项式。当没有指定滞后阶数时，MA 是滞后阶数 1，2，…的移动平均多项式的系数。当用 MALags 指定时，MA 是与 MALags 中的滞后阶数等长的系数向量。默认值：NaN 值的元胞向量。
- MALags——与 MA 系数相关的正整数滞后向量。默认值：与非季节移动平均多项式的阶数相等的整数向量。
- SAR——默认值：元素为 NaN 的元胞向量。季节性自回归模型多项式对应的系数。当没有指定 SARLags 时，AR 是滞后 1，2，…阶数的多项式系数，当指定滞后阶数 SARLags 时，AR 是与滞后数等长的系数向量。
- SARLags——与 SAR 系数相关的正整数滞后向量。默认值为与季节性自回归多项式阶数相等的整数向量。
- SMA——非季节性移动平均模型的可以多项式。当没有指定滞后阶数时，SMA 是滞后阶数 1，2，…的移动平均多项式的系数。当用 SMALags 指定时，SMA 是与 SMALags 中的滞后阶数等长的系数向量。默认值：NaN 值的元胞向量。
- SMALags——与 SMA 系数相关的正整数滞后向量。默认值：与季节移动平均

多项式的阶数相等的整数向量。

- Seasonality——线性时间序列模型中季节差分滞后算子多项式程度的非负整数。默认值：0（无季节性）。
- Variance——模型的方差或者条件方差。默认值：NaN。
- Description——描述模型的字符串标量或字符向量。默认情况下，此参数描述模型的参数形式，例如“ARIMA（1,1,1）模型（高斯分布）”。

（1）使用 arima 函数构建一个已知系数的 AR（3）模型

$X_t=0.06+0.7X_{t-1}+0.2X_{t-2}-0.1X_{t-3}+\varepsilon_t$，其中 ε_t 是均值为 0 方差为 0.01 的高斯分布。

```
Mdl = arima('Constant',0.06,'AR',{0.7,0.2,-0.1},...
    'Variance',0.01)

Mdl =

   ARIMA(3,0,0) Model:
   --------------------
   Distribution: Name = 'Gaussian'
             P: 3
             D: 0
             Q: 0
      Constant: 0.06
            AR: {0.7 0.2 -0.1} at Lags [1 2 3]
           SAR: {}
            MA: {}
           SMA: {}
      Variance: 0.01
```

此时也可以修改模型中的参数，比如 Mdl.Constant = NaN 则可改变模型中的参数。

（2）建立季节性 ARIMA 模型

建立一个没有常数项的季节性 MA 模型

$$X_t=\varepsilon_t+b_1\varepsilon_{t-1}+b_2\varepsilon_{t-12}+b_{12}\varepsilon_{t-12}$$

```
>> Mdl = arima('Constant',0,'MALags',[1,2,12])
```

```
Mdl =

    ARIMA(0,0,12) Model:
    --------------------
    Distribution: Name = 'Gaussian'
             P: 0
             D: 0
             Q: 12
        Constant: 0
              AR: {}
             SAR: {}
              MA: {NaN NaN NaN} at Lags [1 2 12]
             SMA: {}
        Variance: NaN
```

（3）用条件方差模型 GARCH 来设定 ARIMA（具体内容将在第 9 章介绍）

```
>> Mdl = arima(1,0,1);
>> Mdl.Variance = garch(1,1)

Mdl =

    ARIMA(1,0,1) Model:
    --------------------
    Distribution: Name = 'Gaussian'
             P: 1
             D: 0
             Q: 1
        Constant: NaN
              AR: {NaN} at Lags [1]
             SAR: {}
              MA: {NaN} at Lags [1]
             SMA: {}
   Variance: [GARCH(1,1) Model]
```

6.2.2 估计 ARIMA 模型的参数

估计 ARIMA 模型的参数使用函数 estimate，使用方法如下：

EstMdl = estimate(Mdl,y)

[EstMdl,EstParamCov,logL,info] = estimate(Mdl,y)

[EstMdl,EstParamCov,logL,info] = estimate(Mdl,y,Name,Value)

（1）输入变量含义

- Mdl——ARIMA 模型变量。
- y——用来估计参数的时间序列数据。

（2）可选择的输入变量含义

- AR0——非季节性 ARIMA 模型回归系数的初始估计，AR0 中的系数数量必须等于非季节性自回归多项式 ARLags 中与非零系数相关的滞后数。默认情况下，使用标准时间序列技术推导初始值。
- Beta0——回归系数中回归分量的初始估计。
- Constant0——模型中常数项的初始估计。
- Display——命令窗口显示输出选项。
  - 'diagnostics'——优化诊断。
  - 'full'——极大似然参数估计，标准误，*t* 统计量，迭代最优信息，优化诊断。
  - 'iter'——迭代最优信息。
  - 'off'——在命令窗无显示。
  - 'params'——极大似然参数估计，标准误，*t* 统计量。默认值为：'params'。
- DoF0——*t* 分布自由度的初始估计。
- E0——前样本改进量。
- MA0——非季节性移动平均模型系数初始估计。

- Options——最优化选项，运用 Optimization Toolbox。
- SAR0——季节性自回归模型系数初始估计。
- SMA0——季节性滑动均值模型系数初始估计。
- V0——前样本条件方差。
- Variance0——方差初始估计。
- x——外生预测指标。
- Y0——样本前响应数据，为 ARIMA 模型提供初始值。

（3）输出变量含义如下

- EstMdl——模型的估计参数，返回一个 ARIMA 模型。
- EstParamCov——极大似然参数估计的方差-协方差矩阵，包括常数项，季节、非季节性 AR 系数，季节、非季节性 MA 系数，回归系数，方差及自由度。
- logL——最优化极大似然目标方程的值。
- info——总结
  - exitflag——最优化退出标志。
  - options——最优化选项控制因素。
  - x——最终参数估计值。
  - x0——初始参数估计值。

6.3 ARIMA 模型的应用实例

利用 ARIMA 模型可以对股票指数进行建模，用于研判指数的未来走势。我们选用上证 50 指数的数据进行 ARIMA 模型建模，可以用如下脚本实现：

（1）读取指数数据

```
clc, clear all, close all
```

```
[Index,date]=xlsread('Index50','Sheet1','A2:B469');
N = length(Index);
a = zeros(N,1);
for i = 1:N
   a(i) = datenum(cell2mat(date(i)));
end
a = datestr(a);
a = datevec(a);
Dates = datetime(a);
Returns = tick2ret(Index); %指数转收益率
```

（2）原始数据可视化

```
figure(1)
subplot(2,1,1)
plot(Dates,Index);
title('上证 50 股票价格指数')
ylabel('指数')

subplot(2,1,2)
plot(Dates(1:end-1),Returns);
title('上证 50 股票价格指数收益率')
ylabel('收益率')
```

图 6-1 展示了上证 50 指数及指数收益率的数据图像，可以看到收益率数据是相对平稳的，而指数数据有趋势性。

（3）平稳性检验

我们使用 pp 检验来检验数据的平稳性，该检验的具体内容我们将在后面的章节详细给出。

```
disp('使用 PP 检验，如果不能拒绝原假设则说明指数序列存在单位根')
[hp,hpValue,stat,cValue,reg] = pptest(Index,'model','TS')
if hp == 0 %存在单位根则做一阶差分
   diffIndex = diff(Index);
   [hp,hpValue,stat,cValue,reg] = pptest(diffIndex,'model','TS')
end
    [hp,hpValue,stat,cValue,reg] = adftest(diffIndex,'model','TS')
```

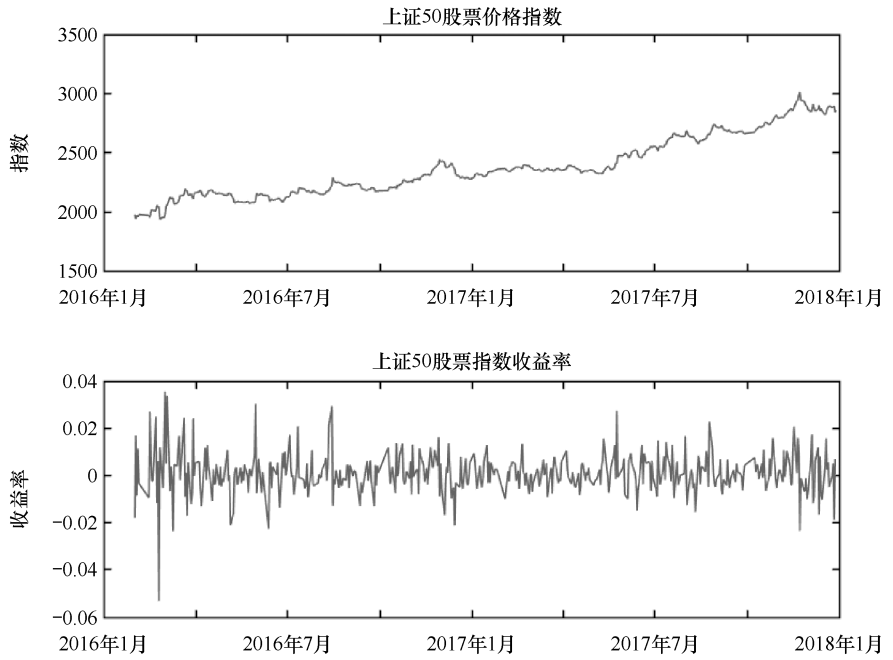


图 6-1　上证 50 指数及指数收益率

首先对原数据进行 pp 检验，如果不能拒绝原假设则说明指数序列存在单位根，结果为：

```
hp =

  logical

   0


hpValue =

    0.2226
```

如上，可以看出对原指数进行 pp 检验，hp=0，说明数据不平稳。因此需要对数据进行一阶差分处理，再对差分数据进行 pp 检验，结果如下：

```
hp =
```

```
  logical

   1


hpValue =

   1.0000e-03
```

对差分数据进行检验，hp=1，说明差分后的数据平稳，我们同时也可以使用 adf 检验，结果如下：

```
hp =

  logical

   1


hpValue =

   1.0000e-03
```

如上，也同样可以得到差分数据平稳的结论。

（4）自相关及偏自相关函数图

```
%原指数图片
figure(2)
subplot(2,1,1)
autocorr(Index);
title('指数的自相关图像')
subplot(2,1,2)
parcorr(Index);
title('指数的偏自相关图像')
%差分后指数图片
figure(3)
```

```
subplot(2,1,1)
autocorr(diffIndex);
title('指数一阶差分后的自相关图像')
subplot(2,1,2)
parcorr(diffIndex);
title('指数一阶差分后的偏自相关图像')
```

由图 6-2 可以看出该数据适用于 ARIMA 模型，同样绘制一阶差分后的图像得到图 6-3。

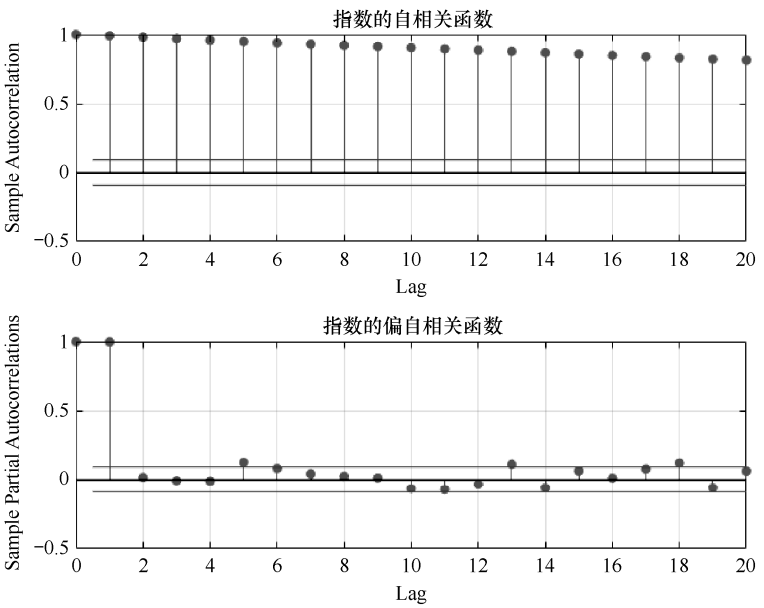


图 6-2　指数自相关及偏自相关函数图

（5）AIC 定阶

```
maxLags = 4;
AICSet = zeros(maxLags, maxLags);
%此处也可以使用 parfor，运用并行减少时间
for i = 1:maxLags
for j = 1:maxLags
      mdl = arima('ARLags',[1:i],'MALags',[1:j]);
      [EstMdl, EstParamCov, logL, info] = estimate(mdl, diffIndex,
      'display', 'off');
```

```
            AICSet(i, j) = aicbic(logL, length(info.X));

        end
        end

        % 画热度图来表示 AIC 数值的分布
        figure(4)
        heatmap(AICSet/1000);
        xlabel('MA Lags')
        ylabel('AR Lags')
        title('Akaike information criteria')

        [OptimalARLags, OptimalMALags] = find(AICSet==min(min(AICSet)));
        title(['Optimal AR and MA Lags are (' num2str(OptimalARLags) ','
        num2str(OptimalMALags) ')'])
```

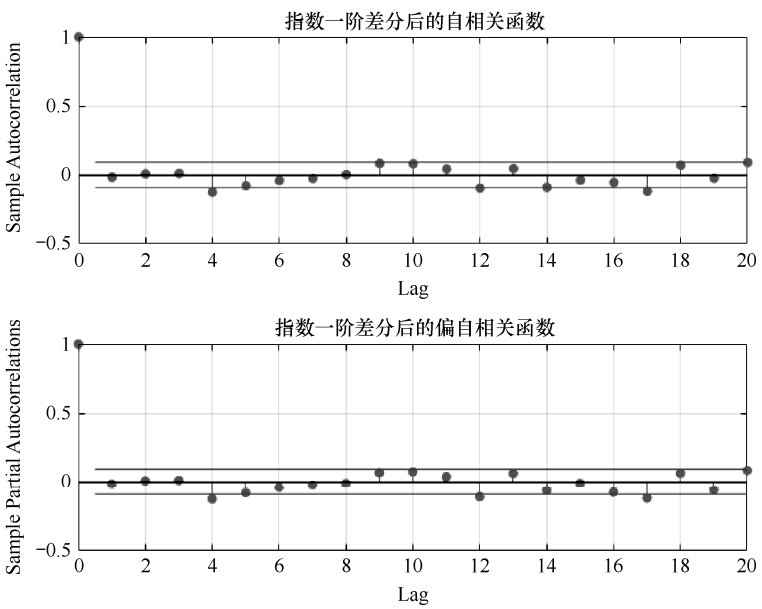


图 6-3　指数一阶差分后的自相关及偏自相关函数图

我们使用 AIC 准则进行模型的定阶，绘制 AIC 准则的热度图如图 6-4 所示，选择其数值最小的可知最优为 ARIMA（4,1,4）。

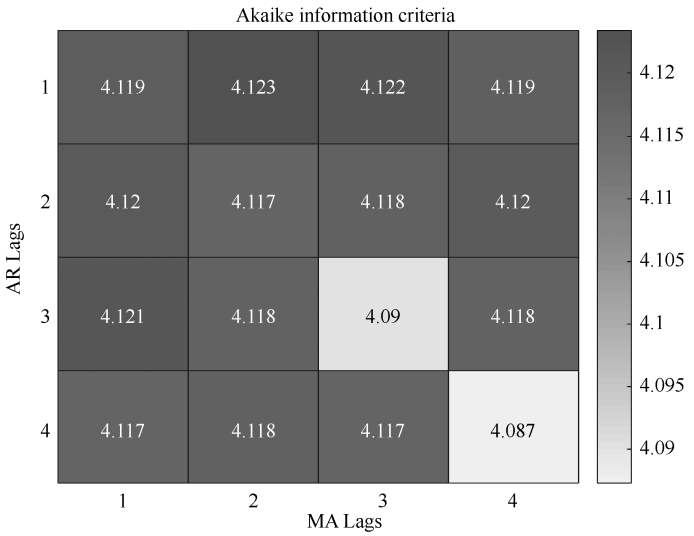


图 6-4　AIC 准则定阶分布图

（6）建立模型

```
mdl = arima( OptimalARLags,1, OptimalMALags);
fit = estimate(mdl, Index);
```

建立 ARIMA（4,1,4）模型如下：

```
ARIMA(4,1,4) Model:
   --------------------
   Conditional Probability Distribution: Gaussian

                            Standard          t
    Parameter       Value       Error     Statistic
   -----------   -----------   ------------   -----------
    Constant       2.29613       1.0227       2.24517
      AR{1}      -0.344011    0.0932337      -3.68978
      AR{2}       1.00797     0.119735       8.41829
      AR{3}      -0.103552    0.0974177      -1.06297
      AR{4}      -0.784864    0.0800389      -9.80603
      MA{1}       0.402167     0.111618       3.60307
      MA{2}       -1.02112     0.139701      -7.30932
      MA{3}     -0.0315672     0.107138     -0.294642
      MA{4}       0.740718    0.0988979       7.48972
    Variance       357.476       17.678       20.2215
```

（7）模型检验

```
adjdiffIndex = diffIndex - mean(diffIndex);
disp(['检验残差是否存在相关性']);
[hLBQ, pLBQ] = lbqtest(adjdiffIndex, 'Lags', 1:4, 'alpha', 0.05)
```

检验残差是否存在相关性得到如下结果：

```
hLBQ =

  1×4 logical array

   0   0   0   0


pLBQ =

  0.7184    0.9256    0.9764    0.1111
```

可以看到模型的残差不具有相关性。

6.4 小结

对于非平稳时间序列，我们可以采用自回归积分滑动平均模型——ARIMA 模型进行建模，该模型的实质是将差分运算与 ARMA 模型结合在一起。非平稳数据可以通过适当阶数的差分变得平稳，当数据平稳后即可用 ARMA 模型进行拟合。在 MATLAB 中可以使用 arima 函数进行建模，该函数可以创建 MA 模型、AR 模型、ARMA 模型、ARIMA 模型及季节性时间序列模型，设定完模型参数后，我们使用函数 estimate 对设定好参数的 arima 模型进行参数估计，利用这两个函数能够在 MATLAB 中便捷地建立 ARIMA 模型。在 6.3 节中我们使用上证 50 指数的数据建立了 ARIMA 模型。

参考文献

[1] 王黎明，王连，杨楠. 应用时间序列分析. 上海：复旦大学出版社，2009.
[2] 黄红梅. 应用时间序列分析. 北京：清华大学出版社，2016.

[3] 孙祝岭. 时间序列与多元统计分析. 上海：上海交通大学出版社，2016.

[4] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[5] Ruey S. Tsay, 王远林，王辉，等. 金融时间序列分析（第三版）. 北京：人民邮电出版社，2012.

[6] Ho S L, Xie M(1998). The use of ARIMA models for reliability forecasting and analysis[J]. Computers & Industrial Engineering, 35(1-2):213-216.

[7] Montanari A, Rosso R, Taqqu M S(1997). Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation[J]. Water Resources Research, 33(5):1035-1044.

7 建模及预测

建立一个时间序列模型一般遵循如图 7-1 所示步骤，本章重点介绍其中几步的具体方法。

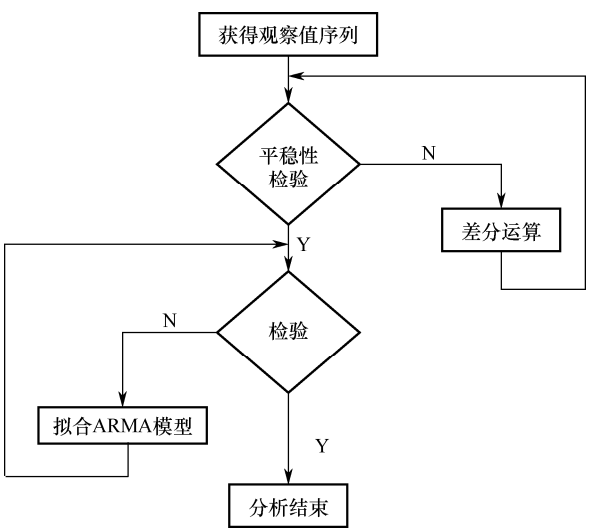


图 7-1　时间序列建模步骤

7.1　平稳性检验方法

假设随机过程 $\{X_t, t=1,2,\cdots,n\}$ 满足 $X_t = X_{t-1} + \varepsilon_t$，其中 ε_t 独立同分布，且 $E(\varepsilon_t)=0$，$D(\varepsilon_t^2)=E(\varepsilon_t^2)=\sigma^2<\infty$，则称 X_t 为随机游走过程。假设随机过程 $\{X_t, t=1,2,\cdots,n\}$ 满足 $X_t = \rho X_{t-1} + \varepsilon_t$，其中 ε_t 为一均值为 0 的平稳序列，当 $\rho=1$ 时，其滞后算子特征多项式存在一个单位根，此时称 X_t 为单位根过程。可见，随机游走过程是单位根过程的特例。当 $\rho<1$ 时，X_t 是一个平稳序列，当 $\rho>1$ 时，X_t 是一个非平

稳序列。

我们已经看到，判断时间序列非平稳性的类型对于后续分析非常重要。在将非平稳序列转换为平稳序列前，首先要确定原序列包含的是确定性趋势还是随机趋势，否则，转换后的序列中可能会出现统计假象。在 Box-Jenkins 理论中，非平稳时间序列都可以包含在 ARIMA（*p*,*d*,*q*）模型中。时间序列分析曾经花了很多时间来寻找能够得到平稳 ARMA 过程的差分阶数 *d*，主要的方法是自相关函数。为此，需要对水平序列以及差分序列的相关函数图进行分析。如果随着滞后阶数的增加自相关系数下降非常缓慢，就说明序列是非平稳的。可以用下面这个经验法则对差分阶数进行判断：在某一差分阶数下，差分序列的自相关系数以很快的速度趋近于零，而且与其他阶数差分序列的方差相比，其方差值最小。

7.1.1 Dickey-Fuller 检验

到目前为止，我们对时间序列平稳性的判断都是非正式的，而且不能区分时间序列的趋势平稳与差分平稳。而单位根检验原则正好可以解决这两方面的问题。单位根检验方法最早由 Wayne A. Fuller（1976），David A. Dickey 和 Wayne A. Fuller（1979,1981）提出。其具体内容如下：

如果 AR（1）过程中有一个单位根，即 $a_1=1$，则得到带漂移项的随机游走过程 $X_t=\delta_1+X_{t-1}+\varepsilon_t$，它可以作为检验的原假设，对应的备择假设为 $|a_1|<1$，代表趋势平稳过程。

如果假设 AR（1）过程均值为零，那么在原假设 $a_1=1$ 下非平稳 AR（1）为无漂移项的随机游走过程，则相应备择假设下的平稳过程变成 $X_t=\rho X_{t-1}+\varepsilon_t$，$|\rho|<1$。

对不同假设的情形进行区分十分重要，因为在原假设下，即使是渐进分布也不再服从标准正态分布，它们还取决于趋势的均值等其他参数的信息。如果我们从最一般的模型开始，三种情形下原假设都是 $\rho=1$，即 AR 部分有一个单位根。可以证明，在这一原假设下，ρ 的最小二乘估计量有向下的偏误，而且呈左偏分布。因此，即使原假设 $\rho=1$ 为真，我们预期其估计值 $\hat{\rho}$ 也小于 1。相应地，常用的 $\hat{\rho}-1$ 的 t 检验统计量不再服从常规的 *t* 分布。Wayne A. Fuller（1976）通过模拟的方法首次给出了三种情形下 t 统计量的临界值。现在人们一般常用更为精确的 James G. MacKinnon

（1991）临界值，也是通过模拟推导出来的。三种情形下的单边假设（备择假设为 $\rho<1$）的临界值分别为：1.94（均值为零）、−2.89（均值不为零）以及−3.46（包含线性趋势）。由于三个临界值的绝对值都远大于 t 统计量的临界值−1.65，因此，若应用传统的 t 分布进行判断，会大大增加过度拒绝原假设的可能性。即使序列包含随机游走过程，也很可能会得出平稳或趋势平稳的错误判断。此外，如果要检验联合原假设 $\alpha=\beta=0$ 且 $\rho=1$，或者 $\beta=0$ 且 $\rho=1$，可以使用 David A. Dickey 和 Wayne A. Fuller（1981）提出的 F 检验，作者同时给出了相应的临界值表。

7.1.2 增广的 Dickey-Fuller 检验

如果 AR（p）过程的阶数 $p>1$，那么一阶自回归过程的单位根检验方法可以很方便地扩展到高阶的情形。对于 AR（p）过程：

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + \varepsilon_t$$

进行参数变换，改写为下面的形式：

$$X_t = \rho X_{t-1} + \theta_1 \Delta X_{t-1} + \theta_2 \Delta X_{t-2} + \cdots + \theta_{p-1} \Delta X_{t-p+1} + \varepsilon_t$$

其中：

$$\rho = \theta_0 = \sum_{j=1}^{p} a_j,\quad \theta_i = -\sum_{j=i+1}^{p} a_j\ ,\quad i = 1,2,3,\cdots,p-1$$

若这一 AR（p）过程存在单位根，则有 $1-a_1-a_2-\cdots-a_p=0$ 或 $\rho=1$。前面所讨论的关于 AR（1）过程的各种假设在这里同样适用，且检验统计量的渐进分布是相同的。我们可以使用与 AR（1）过程相同的临界值。若 AR（p）包含确定性趋势，那么方程在 ADF（Augmented Dickey-Fuller）检验中的表达式为：

$$\Delta X_t = \alpha + \beta t + (\rho-1) X_{t-1} + \theta_1 \Delta X_{t-1} + \cdots + \theta_k \Delta X_{t-k} + \varepsilon_t$$

其中 k 值的选择要以保证残差为纯随机过程为原则。

Pierre Perron（1998）指出，如果数据生成过程是趋势平稳的，但是检验式中未包含时间趋势项，那么此时单位根检验的效力将逐渐丧失，即拒绝随机游走原假设的机会降低，尤其在有限样本下几乎从不拒绝原假设。由此可见，单位根检验的有效性主要取决于刻画数据的模型是否恰当。如果数据表现出确定性趋势，只有当原假设

$H_0:\beta=0$ 不能被拒绝时，才在检验的方程中使用不包含时间趋势的简化形式。对于常数项的处理也是如此。

为此，Pierre Perron（1988）建议用下面的方法进行单位根检验：

$$\Delta X_t=\alpha+\beta(t-T/2)+(\rho-1)X_{t-1}+\sum_{i=1}^{k}\theta_i\Delta X_{t-i}+\varepsilon_t$$

T 代表样本容量，为了不影响常数项的估计，对趋势变量进行了中心化处理。应用 Dickey-Fuller 的 t 统计量对原假设 $H_0:\rho=1$ 进行检验；对应的备择假设为 X_t 是趋势平稳的。也可以用 F 检验对联合原假设 $H_0:(\alpha,\beta,\rho)=(\alpha,0,1)$ 进行检验。如果拒绝原假设，则可以假定存在确定性趋势，并进一步检验 $H_0:\beta=0$。如果两个检验都不能拒绝原假设，第二步在以下形式的方程中进行检验。

$$\Delta X_t=\alpha+(\rho-1)X_{t-1}+\sum_{i=1}^{k}\theta_i\Delta X_{t-i}+\varepsilon_t$$

同样，对原假设 $H_0:\rho=1$ 进行 t 检验，即检验是否存在单位根。此时，备择假设为 X_t 是均值非零的平稳 AR 过程。

若还要检验常数项是否为零，可以对 $H_0:(\alpha,\rho)=(0,1)$ 进行 F 检验。如果不能拒绝原假设，则在以下模型中:

$$\Delta X_t=(\rho-1)X_{t-1}+\sum_{i=1}^{k}\theta_i\Delta X_{t-i}+\varepsilon_t$$

对 $H_0:\rho=1$ 进行检验。

在所有情形下，都可以用 James G. MacKinnon（1991）中的 t 检验临界值，以及 David A. Dickey 和 Wayne A. Fuller（1981）中的 F 检验临界值。

7.1.3 Phillips-Perron 检验

针对误差项可能存在的自相关或异方差，Peter C.B. Phillips 和 Pierre Perron（1988）提出了另一种平稳性检验方法。Phillips-Perron 检验的方法是：利用残差自相关中的信息对系数 $\hat{\rho}$ 的长期方差进行非参数估计，并将此估计值用于对原假设 ρ=1 的检验值进行调整。

两位作者建议用下式估计残差的长期调整方差：

$$s_{Tm}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{u}_t^2 + \frac{2}{T}\sum_{i=1}^{T-1}\left(w_{im}\sum_{t=i+1}^{T}\hat{u}_t\hat{u}_{t-i} \right)$$

其中 $\hat{u}$ 为方程的最小二乘残差； m 为截断参数，代表引入自协方差的最大滞后阶数。当样本容量 T 趋于无穷时， m 也随之增加，但增加的速度小于 T 。 w_{im} 为权重，可以保证长期方差估计量的一致性和非负性。对于 w_{im} 的确定，Pierre Perron（1988）建议采用 Maurice Stevenson Bartlett（1948）提出的权重：

$$w_{im} = \begin{cases} 1-\dfrac{i}{m+1}, i = 1,\cdots,m \\ 0, i > m \end{cases}$$

利用这一调整后的方差，我们可以对带有时间趋势以及常数项的模型进行 F 检验，原假设为 $\mathrm{H}_0 : (\alpha,\beta,\rho) = (\alpha,0,1)$ ，检验统计量为：

$$\tilde{F}_{Tr} = \frac{s}{s_{Tm}}\hat{F}_{Tr} - \frac{(s_{Tm}^2 - s^2)}{2s_{Tm}^2}\left[T(\hat{\rho}-1) - \frac{T^6(s_{Tm}^2 - s^2)}{48|X'X|} \right]$$

式中， s 是估计的回归标准差； X 是前定变量矩阵；前三个列向量分别由 1， X_{t-1} 和 t 构成， $X = [1\ X_{t-1}\ t\ ]$ 。

给定以上的原假设， $\hat{F}_{Tr}$ 是传统的 F 统计量。在对包含趋势项的模型进行原假设为 $\mathrm{H}_0 : \rho = 1$ 的检验时，一般采用 t 检验，建议采用下面这个调整后的统计量：

$$\tilde{t}_{Tr} = \frac{s}{s_{Tm}}\hat{t}_{Tr} - \frac{(s_{Tm}^2 - s^2)T^3}{4s_{Tm}\sqrt{3|X'X|}}$$

其中 $\hat{t}_{Tr}$ 代表常规的 t 统计量。

如果检验不能拒绝相应的原假设，意味着可能不存在确定性趋势。在这种情况下，可以对更强的原假设 $\mathrm{H}_0 : (\alpha,\beta,\rho) = (0,0,1)$ 进行检验，检验统计量如下：

$$\tilde{F}_{Tr} = \frac{s}{s_{Tm}}\hat{F}_{Tr} - \frac{(s_{Tm}^2 - s^2)}{3s_{Tm}^2}\left[T(\hat{\rho}-1) - \frac{T^6(s_{Tm}^2 - s^2)}{48|X'X|} \right]$$

在数据无确定性趋势的假定下，应用统计量 $\tilde{F}_{\mu}$ 来联合检验假设。

$$\tilde{F}_{\mu}=\frac{s}{s_{Tm}}\hat{F}_{\mu}-\frac{(s_{Tm}^2-s^2)}{2s_{Tm}^2}\left[T(\hat{\rho}-1)-\frac{T^2(s_{Tm}^2-s^2)}{4\sum_{i=1}^{T}(y_t-\bar{y})^2}\right]$$

联合假设为 $\mathrm{H}_0:(\alpha,\rho)=(0,1)$。其中 $\hat{F}_{\mu}$ 是这一原假设下的常规 F 统计量。如果不能拒绝原假设，可以在无确定性成分的模型中检验 $\mathrm{H}_0:\rho=1$，检验统计量为:

$$\tilde{t}_{\rho}=\frac{s}{s_{Tm}}\hat{t}_{\rho}-\frac{0.5(s_{Tm}^2-s^2)T}{s_{Tm}\sqrt{\sum_{t=2}^{T}y_{t-1}^2}}$$

也就是说，我们检验时间序列是否包含一个无漂移项的随机游走过程。如果该假设被拒绝，那么继续应用统计量 $\tilde{t}_{\mu}$ 来检验是否存在带有漂移项的随机游走过程:

$$\tilde{t}_{\mu}=\frac{s}{s_{Tm}}\hat{t}_{\mu}-\frac{0.5(s_{Tm}^2-s^2)T}{s_{Tm}\sqrt{\sum_{t=2}^{T}(y_t-\bar{y})^2}}$$

$\hat{t}_{\mu}$ 和 $\hat{t}_{\rho}$ 同样是在这一原假设下的常规 t 统计量。在以上所有情形下，都可以应用 James G. MacKinnon（1991）给出的 t 检验临界值以及 David A. Dickey 和 Wayne A. Fuller（1981，1063）给出的 F 检验临界值。

7.2 AIC 准则定阶

对于模型的阶数，我们需要用更严谨的方式来进行判定，因此我们介绍 AIC 准则定阶法。

AIC 准则（A-Information Criterion 最小信息准则）首先由日本学者赤池（Akaike）提出，也称为赤池信息准则。它适用于 AR、MA、ARMA 三类模型的定阶。

设 $\{X_t:1\leqslant t\leqslant N\}$ 为一随机时间序列，对其拟合 ARMA（p，q）模型，用极大似然方法估计模型的参数，L 是模型的极大似然值，AIC 准则函数定义如下：

$$\begin{aligned}\mathrm{AIC}(p,q) &= -2\ln[L]+2r \\ &\approx N\ln(\hat{\sigma}_a^2)+2r+C\end{aligned}$$

其中，$r=p+q$ 为模型的独立参数个数；$\hat{\sigma}_a^2$ 是残差方差的极大似然估计；C 是常数。实际中也常用如下定义的 AIC 准则函数（用样本大小 N 标准化）：

$$\mathrm{AIC}(p,q)=\ln[\hat{\sigma}_a^2]+2r/N$$

AIC 准则函数中的 $\hat{\sigma}_a^2$ 应是模型残差方差的极大似然估计，但由于极大似然估计求解困难，实际中也可用矩估计或最小二乘估计所得残差方差近似代替。

可以看出，AIC 准则函数由两部分构成：第一部分反映模型拟合的好坏；第二部分表明模型参数的多少。我们总是希望模型拟合得越精确越好，并且模型中的参数尽可能少一些。AIC 准则函数就是将这两个目标进行适当的综合。当模型阶数增高时，AIC 准则函数中的第一项一般是下降的。对给定的观察数据个数 N，式中的第二项随模型阶数而增长。当逐次增加模型阶数对数据进行拟合时，AIC 的值是有下降趋势的，因为这时模型的残差方差下降较快，起决定性作用的是第一部分。当达到某一阶数时，AIC 值达到极小。随后，随着模型阶数继续增高，残差方差改进甚微，于是第二部分（模型阶数）起关键性作用，AIC 的值随模型阶数而增长。对事先给定的最高阶数 M（N），如果：

$$\mathrm{AIC}(p_0,q_0)=\min_{1\leqslant p,q\leqslant M(N)}\mathrm{AIC}(p,q)$$

我们便取 n_0 和 m_0 为最佳模型阶数；拟合模型的最高阶数 $M(N)$通常取为 $[N/3]\sim[2N/3]$ 之间的某个整数。

7.3 模型的检验

完成模型的稳定性检验、定阶和参数估计后，剩下的问题就是判断这个模型用于描述时间序列是否恰当，也就是模型的适应性检验。所谓模型的适应性是指一个时间序列模型解释系统动态性（即数据序列的相关性）的程度。一个时间序列的适合模型应该完全或基本上解释了系统的动态性，从而模型中残差序列应该是白噪声序列。模

型的适应性检验实质上就是检验$\{\varepsilon_t\}$序列是否为白噪声序列。一个模型是否显著有效主要看它提取的信息是否充分。

一个好的拟合模型应该能够提取观察值序列中几乎所有的样本相关信息，换言之，拟合残差项中将不再蕴含任何相关信息，即残差序列应该为白噪声序列。这样的模型称为显著有效模型。反之，如果残差序列为非白噪声序列，那就意味着残差序列中还残留着相关信息未被提取，这就说明拟合模型不够有效，通常需要选择其他模型重新拟合。

所以模型的显著性检验即为残差序列的白噪声检验。原假设和备择假设分别为：

$$H_0:\varepsilon_1=\varepsilon_2=\cdots=\varepsilon_m=0,\forall m\geqslant 1$$

$$H_1:\text{至少存在某个}\varepsilon_k\neq 0,\forall m\geqslant 1,k\leqslant m$$

检验统计量为 LB(Ljung-Box) 检验统计量：

$$LB=n(n+2)\sum_{k=1}^{m}\left(\frac{\hat{\varepsilon}_k^2}{n-k}\right)\sim\chi^2(m),\forall m>0$$

如果拒绝原假设，就说明残差序列中还残留着相关信息，拟合模型不显著。如果不能拒绝原假设，就认为拟合模型显著有效。

7.4 ADF 检验方法的 MATLAB 实现

7.4.1 ADF 检验

ADF 检验使用的函数是 adftest，使用方法如下：

h = adftest (Y)

h = adftest (Y,Name,Value)

[h,pValue] = adftest (___)

[h,pValue,stat,cValue,reg] = adftest (___)

（1）函数的输入变量含义

单变量时间序列数据。最后一个元素是时间最近的观察值，NaN 表示被移除的缺失值。

（2）可选择的输入参数含义

- lags——默认值为 0。非负整数标量或向量，用 Newey-West 估计的长期方差时自协方差的滞后阶数。'lags',0:2 表示分别用 0，1，2 三种滞后差分项做检验。
- model——默认值为'AR'，字符向量，例如'AR'是由字符向量组成的元胞向量，表示模型变量，有如下选择：
  - 'AR'（自回归）

    检验的原假设为：

    $$X_t = X_{t-1} + a_1 \Delta X_{t-1} + \cdots + a_p \Delta X_{t-p} + \varepsilon$$

    备择假设为：

    $$X_t = bX_{t-1} + a_1 \Delta X_{t-1} + \cdots + a_p \Delta X_{t-p} + \varepsilon, b < 1$$

  - 'ARD'（带漂移项的自回归）

    检验的原假设为：

    $$X_t = X_{t-1} + a_1 \Delta X_{t-1} + \cdots + a_p \Delta X_{t-p} + \varepsilon$$

    备择假设为：

    $$X_t = c + bX_{t-1} + a_1 \Delta X_{t-1} + \cdots + a_p \Delta X_{t-p} + \varepsilon, b < 1$$

  - 'TS'（趋势稳定）

    趋势稳定模型变量，检验的原假设是：

    $$X_t = c + X_{t-1} + a_1 \Delta X_{t-1} + \cdots + a_p \Delta X_{t-p} + \varepsilon$$

    备择假设为：

$$X_t = c + dt + bX_{t-1} + a_1\Delta X_{t-1} + \cdots + a_p\Delta X_{t-p} + \varepsilon,\ b < 1$$

带有常数项系数 c 及趋势系数 b。

- test——默认't1'，字符向量，表示检验统计量。
  - 't1'——标准 t 统计量。
  - 't2'——滞后调整，非标准 t 统计量。
  - 'F'——F 统计量。
- alpha——默认 0.01.表示检验的显著性水平，取值从 0.001 到 0.999。使用方法如 'alpha',0.01。

（3）输出变量的含义

- h——测试的布尔决策向量，长度等于测试的次数。值等于 1 则表示拒绝有单位根的原假设。值等于 0 则表示不能拒绝原假设，模型有单位根。
- pValue——检验统计量的 p 值向量，长度等于测试次数。如果统计量为't1'和't2'，则 p 值为左尾概率；如果统计量为 'F'，则 p 值为右尾概率。

  当检验统计量超出列表的临界值时，pptest 的 p 值返回最大值（0.999）或最小值（0.001）。
- stat——检验统计量，长度等于检验的次数。在备择假设中使用 OLS 方法估计系数来计算统计量。
- cValue——检验的关键值向量，长度等于检验的数量。值取左尾概率。
- reg——OLS 估计系数的不同回归结构，记录数等于测试数，有如下内容：
  - num——时间序列长度。
  - size——有效的样本量，根据滞后数调整。
  - names——回归系数名。
  - coeff——估计的系数值。

- se——估计系数的标准差。
- cov——估计系数的方差矩阵。
- tStats——t 统计量的系数及概率 p 值。
- FStat——F 统计量及 p 值。
- yMu——输入序列调整滞后数的均值。
- ySigma——输入序列调整滞后数的标准差。
- yHat——输入序列调整滞后数的适应值。
- res——回归残差项。
- DWStat——Durbin-Watson 统计量。
- SSR——回归平方和。
- SSE——误差平方和。
- SST——总平方和。
- MSE——均方误差。
- RMSE——回归的标准误。
- RSq——R^2 统计量。
- aRSq——调整后的 R^2 统计量。
- LL——高斯似然数。
- AIC——AIC 值。
- BIC——BIC 值。
- HQC——Hannan-Quinn 信息准则。

7.4.2 pp 检验

pp 检验使用的函数是 pptest，使用方法如下：

[h,pValue,stat,cValue,reg] = pptest(y)

[h,pValue,stat,cValue,reg] = pptest(y,'ParameterName',ParameterValue,...)

（1）输入参数的含义

- y——时间序列数据向量。最后一个元素是时间最近的观察值，NaN 表示被移除的缺失值。

（2）可选择的输入参数含义

- lags——默认值为 0。非负整数标量或向量，用 Newey-West 估计的长期方差时自协方差的滞后阶数。
- model——默认值为'AR'，字符向量，例如'AR'或者由字符向量组成的元胞向量，表示模型变量，有如下选择：
  - 'AR'（自回归）

    pptest 检验的原假设为：

    $X_t = X_{t-1} + \varepsilon$

    备择假设为：

    $X_t = aX_{t-1} + \varepsilon$，$a < 1$
  - 'ARD'（带漂移项的自回归）

    'AR'模式中原假设和备择假设都增加一个常数项系数：

    $X_t = c + aX_{t-1} + \varepsilon$，$a < 1$
  - 'TS'（趋势稳定）

    pptest 检验的原假设为：

    $X_t = c + X_{t-1} + \varepsilon$

    备择假设为：

    $X_t = c + bt + aX_{t-1} + \varepsilon$，$a < 1$

带有常数项系数 c 及趋势系数 b。

- test——默认't1'，字符向量，表示检验统计量。

  't1'，pptest 检验标准 t 统计量。

  't2'，pptest 检验修正的非标准 t 统计量。

- alpha——默认 0.05，表示检验的显著性水平，取值从 0.001 到 0.999。

（3）输出变量的含义

- h——测试的布尔决策向量，长度等于测试的次数。值等于 1 则表示拒绝有单位根的原假设，模型没有单位根。值等于 0 则表示不能拒绝原假设，模型有单位根。

- pValue——检验统计量的 p 值向量，长度等于测试次数，p 值是左尾概率。

  当检验统计量超出列表的临界值时，pptest 的 p 值返回最大值（0.999）或最小值（0.001）。

- stat——检验统计量，长度等于检验的次数。在备择假设中使用 OLS 方法估计系数来计算统计量。

- cValue——检验的关键值向量，长度等于检验的数量。值取左尾概率。

- reg——OLS 估计系数的不同回归结构，记录数等于测试数，有如下内容：

  - num——时间序列长度。
  - size——有效的样本量，根据滞后数调整。
  - names——回归系数名。
  - coeff——估计的系数值。
  - se——估计系数的标准差。
  - cov——估计系数的方差矩阵。
  - tStats——t 统计量的系数及概率 p 值。

- FStat——F 统计量及 p 值。
- yMu——输入序列调整滞后数的均值。
- ySigma——输入序列调整滞后数的标准差。
- yHat——输入序列调整滞后数的适应值。
- res——回归残差项。
- autoCov——估计残差的自协方差。
- NWEst——Newey-West 估计量。
- DWStat——Durbin-Watson 统计量。
- SSR——回归平方和。
- SSE——误差平方和。
- SST——总平方和。
- MSE——均方误差。
- RMSE——回归的标准误差。
- RSq——R^2 统计量。
- aRSq——调整后的 R^2 统计量。
- LL——高斯似然数。
- AIC——AIC 值。
- BIC——BIC 值。
- HQC——Hannan-Quinn 信息准则。

7.4.3 AIC 准则值

AIC 准则定阶使用的函数是 aicbic，该函数可以计算模型 AIC 及 BIC 的值，使用方法如下：

aic = aicbic(logL,numParam) 返回估计模型的 AIC 值

[aic,bic] = aicbic(logL,numParam,numObs) 返回估计模型的 AIC 或 BIC 值

（1）输入参数的含义

- logL——最优极大似然值。
- numParam——与 logL 相应的估计参数值。
- numObs——样本大小。

（2）输出参数的含义

- aic——AIC 值。
- bic——BIC 值。

我们使用一个特定模型来解释如何使用该函数，比较模型的 AIC、BIC 准则值。特定模型为：$X_t = -3 + 0.3X_{t-1} + 0.4X_{t-2} + \varepsilon_t$，建立符合该模型的数据，程序如下：

```
rng(1);
T = 100; % Sample size
DGP = arima('Constant',-3,'AR',[0.3, 0.4], 'Variance',2)
y = simulate(DGP,T);

DGP =

   ARIMA(2,0,0) Model:
   --------------------
   Distribution: Name = 'Gaussian'
            P: 2
            D: 0
            Q: 0
      Constant: -3
           AR: {0.3 0.4} at Lags [1 2]
          SAR: {}
           MA: {}
          SMA: {}
      Variance: 2
```

定义三种不同的 AR 模型：

```
EstMdl1 = arima('ARLags',1);
EstMdl2 = arima('ARLags',1:2);
EstMdl3 = arima('ARLags',1:3);
```

估计模型参数如下：

```
logL = zeros(3,1);
[~,~,logL(1)] = estimate(EstMdl1,y,'Display','off');
[~,~,logL(2)] = estimate(EstMdl2,y,'Display','off');
[~,~,logL(3)] = estimate(EstMdl3,y,'Display','off');
```

计算模型的 AIC、BIC 准则值为：

```
[aic,bic] = aicbic(logL, [3; 4; 5], T*ones(3,1))
aic =

  372.9984
  358.4456
  358.9418
```

可知 AR（2）模型的准则值最低，为最优模型。

7.4.4 残差自相关性检验

在 MATLAB 中进行 Ljung-Box Q 检验使用的函数是 lbqtest，使用方法如下：

h = lbqtest(res)

h = lbqtest(res,Name,Value)

[h,pValue] = lbqtest(___)

[h,pValue,stat,cValue] = lbqtest(___)

（1）输入变量含义

- res——用于检验的残差序列，丢失数据用 NaN 代替。

（2）可选择输入变量

- lags——检验的滞后项数目，可以是正整数或者正整数向量。默认值：min ([20,

T−1])。

- Alpha——假设检验的显著性水平。默认值：0.05。
- DoF——检验统计量的自由度。默认值：Lags。

（3）输出变量的含义

- h——测试的结果，其长度等于测试的次数。
  - h = 1——表示拒绝残差无自相关的零假设而选择备择假设。
  - h = 0——表示不拒绝残差没有自相关零假设。即说明残差没有自相关性。
- pValue——检验统计量的概率 p 值。
- stat——检验统计量。
- cValue——检验统计量的关键值。

7.5 模型的预测

根据建立好的模型，我们可以方便地预测出时间序列数据的未来值。详细的理论此处不赘述，在 MATLAB 中我们使用 forecast 函数实现对模型未来值的预测，该函数使用方法如下：

[Y,YMSE] = forecast(Mdl,numPeriods)

[Y,YMSE,V] = forecast(Mdl,numPeriods)

[Y,YMSE,V] = forecast(Mdl,numPeriods,Name,Value)

（1）输入参数的含义

- Mdl——ARIMA 模型。
- numPeriods——预测未来值的范围。

（2）可选择的输入变量的含义

- E0——均值为 0，为模型提供初值，至少要 Mdl.Q 行。

- V0——前样本条件方差，为条件方差模型提供初值，如果模型的方差是常数则没必要。
- X0——前样本预测数据，用来指示条件均值模型中是否存在回归分量。至少包含 Mdl.P 的行数。默认为 Mdl.Beta。
- XF——XF 和 X0 具有相同的列数且至少有 numPeriods 行。XF 的第 *i* 行包含 X0 的 *i* 期之前的预测值。
- Y0——前样本响应数据，为模型提供数据初值。

（3）输出变量的含义

- Y——第 *i* 行包含第 *i* 期的条件均值预测值。
- YMSE——Y 的均值平方误。
- v——预测模型的最小均值平方误。

7.6 模型的建立及预测应用实例

用时间序列方法可以对股市指数数据进行建模，预测分析其未来走势，比如使用上证 180 指数的数据进行建模预测分析，可以用如下脚本实现。

（1）读取数据

```
clc, clear all, close all
[Index,date]=xlsread('Index180','Sheet1','A2:B399');
date1 =693960+Index(:,1); %日期数据转化
N = length(Index);
a = zeros(N,1);
for i = 1:N
   a(i) = datenum(cell2mat(date(i)));
end
a = datestr(a);
a = datevec(a);
Dates = datetime(a);
```

```
Returns = tick2ret(Index); %指数转收益率
```

（2）原数据可视化

```
figure(1)
subplot(2,1,1)
plot(Dates,Index);
title('上证 180 指数')
ylabel('指数')

subplot(2,1,2)
plot(Dates(1:end-1),Returns);
title('上证 180 指数收益率')
ylabel('收益率')
```

图 7-2 是上证 180 的指数及收益率图，可以看出上证 180 指数有趋势性。

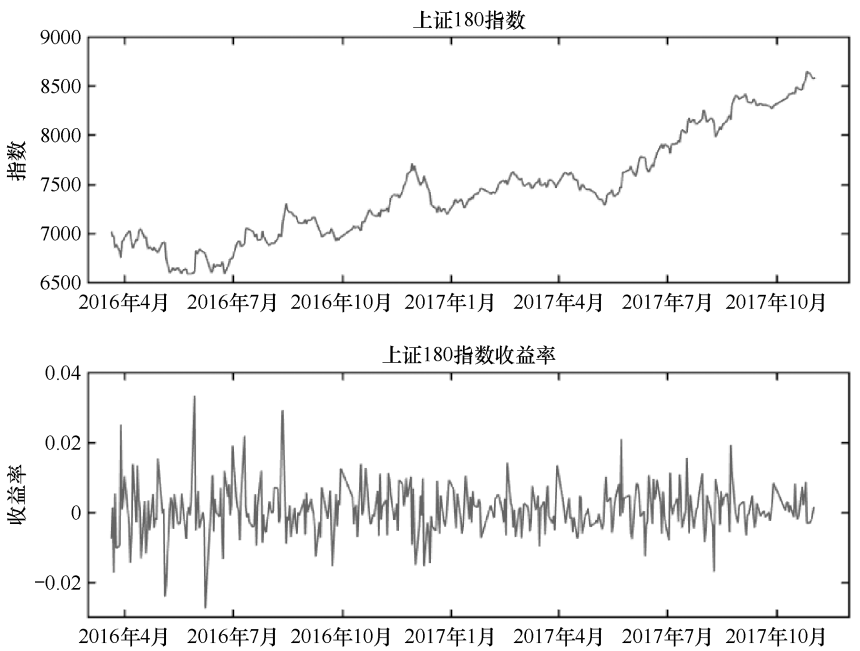


图 7-2　上证 180 指数及收益率图

（3）把数据分为训练集和检验集

因为要对数据进行预测，所以我们将数据分为 70%的训练集及 30%的检验集。

```
rg = round(0.85 * N);
TrainingData = Index(1:trg); % 前 70%
TestData = Index(trg+1:end); % 后 30%
```

（4）平稳性检验

```
disp('使用 PP 检验，如果不能拒绝原假设，则说明指数序列存在单位根')
[hp,hpValue,stat,cValue,reg] = pptest(TrainingData,'model','TS')
if hp == 0 %存在单位根则做一阶差分
    diffIndex = diff(TrainingData);
    [hp,hpValue,stat,cValue,reg] = pptest(diffIndex,'model','TS')
end
   [hp,hpValue,stat,cValue,reg] = adftest(diffIndex,'model','TS')
```

建立模型之前我们要先检验数据的平稳性，因此我们使用 pp 检验来检验训练集数据，如果不能拒绝原假设，则说明指数序列存在单位根。

```
hp =

  logical

   0


hpValue =

    0.0919
```

可以看出对训练集进行 pp 检验，hp=0，说明数据不平稳。因此需要对数据进行一阶差分处理，再对差分数据进行 pp 检验，检验结果如下。

```
hp =

  logical

   1


hpValue =
```

```
    1.0000e-03
```

对差分数据进行检验，hp=1，说明差分后的数据平稳，我们同时也可以使用 ADF 检验，检验结果如下。

```
hp =

  logical

   1


hpValue =

    1.0000e-03
```

ADF 检验同样可以得到差分数据平稳的结论。

（5）自相关及偏自相关图像

```
%原指数图片
figure(2)
subplot(2,1,1)
autocorr(TrainingData);
title('指数的自相关图像')
subplot(2,1,2)
parcorr(TrainingData);
title('指数的偏自相关图像')
%差分后指数图片
figure(3)
subplot(2,1,1)
autocorr(diffIndex);
title('一阶差分后的自相关图像')
subplot(2,1,2)
parcorr(diffIndex);
title('一阶差分后的偏自相关图像')
```

可以通过图 7-3 自相关图像看出该数据适用于 ARIMA 模型，我们绘制一阶差分后的图像如图 7-4 所示。

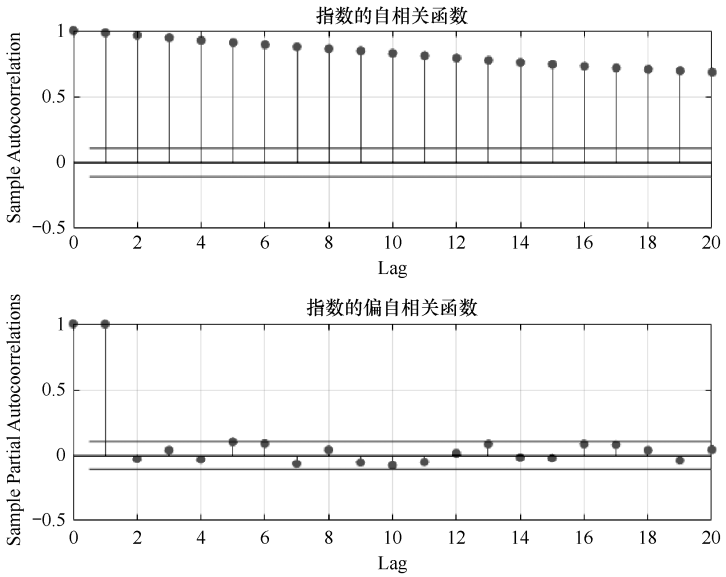


图 7-3　指数自相关及偏自相关函数图

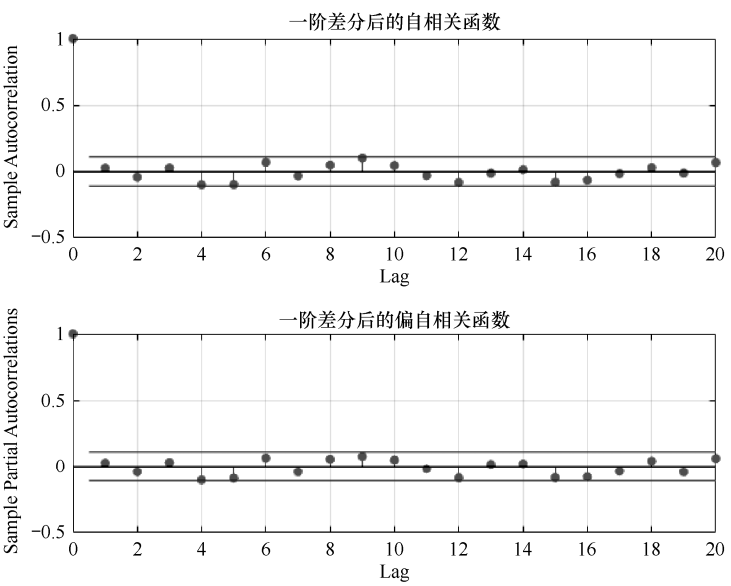


图 7-4　指数一阶差分自相关及偏自相关函数图

（6）AIC 准则定阶

```
maxLags = 4;
AICSet = zeros(maxLags, maxLags);
```

```
parfor i = 1:maxLags
for j = 1:maxLags
       mdl = arima('ARLags',[1:i],'MALags',[1:j]);
       [EstMdl, EstParamCov, logL, info] = estimate(mdl, diffIndex,
       'display', 'off');
       AICSet(i, j) = aicbic(logL, length(info.X));

end
end

% 画热度图来表示 AIC 数值的分布

figure(4)
heatmap(AICSet/1000);
xlabel('MA Lags')
ylabel('AR Lags')
title('Akaike information criteria')

[OptimalARLags, OptimalMALags] = find(AICSet==min(min(AICSet)));
title(['Optimal AR and MA Lags are (' num2str(OptimalARLags) ','
num2str(OptimalMALags) ')'])
```

为使建模更加严谨，我们使用 AIC 准则进行模型的定阶，绘制 AIC 准则的图像如图 7-5 所示，选择其数值最小可知最优模型为 ARIMA（4,1,4）。

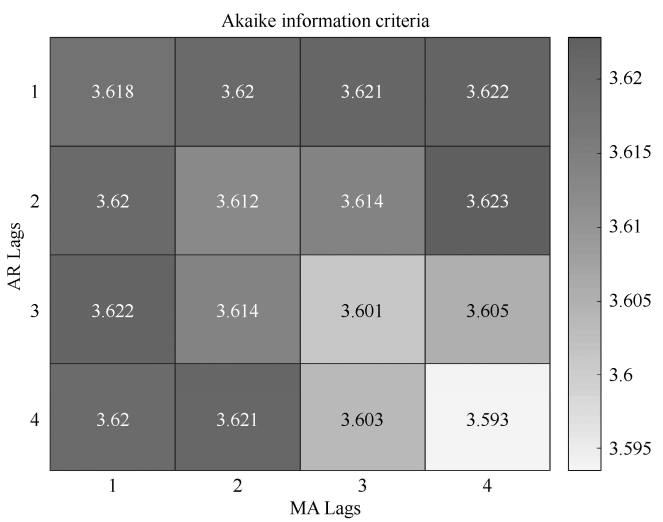


图 7-5　AIC 准则定阶分布图

（7）建模

```
mdl = arima( OptimalARLags,1, OptimalMALags);
fit = estimate(mdl, Index);
```

建立 ARIMA（4,1,4）模型如下：

```
ARIMA(4,1,4) Model:
   --------------------
   Conditional Probability Distribution: Gaussian

                            Standard          t
    Parameter       Value       Error     Statistic
   -----------   -----------   ------------   -----------
    Constant       1.62283      2.41941      0.670756
       AR{1}      0.258448    0.0244403       10.5747
       AR{2}      0.326912     0.034131       9.57815
       AR{3}      0.161484    0.0338029       4.77723
       AR{4}     -0.894298    0.0204882      -43.6493
       MA{1}     -0.240334    0.0321406      -7.47758
       MA{2}     -0.398656    0.0428603       -9.3013
       MA{3}     -0.128176    0.0389652      -3.28949
       MA{4}      0.932089    0.0279378        33.363
    Variance       2276.43      141.663       16.0693
```

（8）模型检验

```
adjdiffIndex = diffIndex - mean(diffIndex);
disp(['检验残差是否存在相关性']);
[hLBQ, pLBQ] = lbqtest(adjdiffIndex, 'Lags', 1:4, 'alpha', 0.05)
```

检验结果如下：

```
hLBQ =

  1×4 logical array

   0   0   0   0


pLBQ =
```

```
    0.6353    0.6516    0.7777    0.3397
```

可以看到残差不具有相关性，因此模型可以信任。

（9）预测未来趋势

```
[Yf,YMSE] = forecast(fit,60,'Y0',TrainingData);
upper = Yf + 1.96*sqrt(YMSE);
lower = Yf - 1.96*sqrt(YMSE);

figure(5)
plot(Index,'b')
hold on
h1 = plot(trg+1:trg+60,Yf,'r','LineWidth',2);
h2 = plot(trg+1:trg+60,upper,'k--','LineWidth',1.5);
plot(trg+1:trg+60,lower,'k--','LineWidth',1.5)
title('95%置信区间')
legend([h1,h2],'Forecast','95% Interval','Location','NorthWest')
hold off
```

我们利用建好的模型进行预测，预测训练集后 60 天的数据，并画出预测数据的95%置信区间如图 7-6 所示。可以看到未来走势在 95%置信区间内。

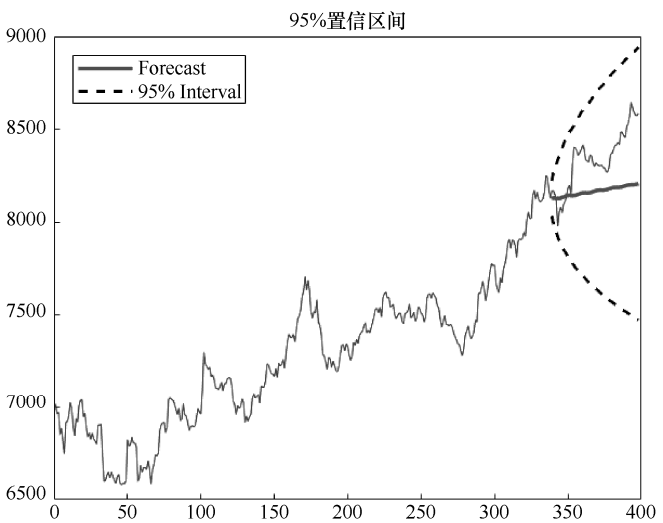


图 7-6　上证 180 指数模型预测

7.7 小结

在介绍了时间序列建模的方法后，本章详细介绍了建模的步骤及利用模型进行预测的方法。首先给出了平稳性检验的方法，主要方法有 Dickey-Fuller 检验，增广的 Dickey-Fuller 检验及 Phillips-Perron 检验，这些检验可以在 MATLAB 中利用函数 adftest 和 pptest 实现。我们在前面介绍过利用自相关函数及偏相关函数图定阶的方法，但是对于模型的阶数，我们需要用更严谨的方法进行判断，因此我们介绍了 AIC 准则定阶法，在 MATLAB 中我们可以利用函数 aicbic 算出模型的 AIC 准则值，从而选择 AIC 值最小的模型作为最优的模型。在完成模型的稳定性检验、定阶、参数估计后，我们还需要判断这个模型用于描述时间序列是否恰当，因此我们介绍了残差自相关性检验 Ljung-Box Q 检验，该检验可以用函数 lbqtest 实现。在 7.4 节我们对这些函数的使用方法进行了详细的介绍。7.5 节我们介绍了如何使用建立的模型进行预测，主要是使用函数 forecast。在 7.6 节我们选用了上证 180 指数数据建立了时间序列模型并用该模型进行了预测。

参考文献

[1] Torres J L, García A, Blas M D, et al(2005). Forecast of hourly average wind speed with ARMA models in Navarre (Spain)[J]. Solar Energy, 79(1):65-77.

[2] Ho S L, Xie M(1998). The use of ARIMA models for reliability forecasting and analysis[J]. Computers & Industrial Engineering, 35(1-2):213-216.

[3] Rojas I, Valenzuela O, Rojas F, et al(2008). Soft-computing techniques and ARMA model for time series prediction[J]. Neurocomputing, 71(4):519-537.

[4] Contreras J, Espinola R, Nogales F J, et al(2003). ARIMA models to predict next-day electricity prices[J]. IEEE Transactions on Power Systems, 18(3):1014-1020.

[5] Kavasseri R G, Seetharaman K(2009). Day-ahead wind speed forecasting using-ARIMA models[J]. Renewable Energy, 34(5):1388-1393.

[6] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[7] Ruey S. Tsay, 王远林，王辉，等. 金融时间序列分析（第三版）. 北京：人民邮电出版社，2012.

[8] Jonathan D.Cryer, Kung-Sik Chan. Time Series Analysis With Applications in R (Second Edition): New York, Springer, 2008.

8 趋势及季节性时间序列建模

在现实生活中产生的非平稳序列通常会显示出非常明显的规律性，比如有显著的趋势或者有固定的变化周期，这种规律性信息通常比较容易提取。人们发现，尽管不同序列的情况千变万化，但是序列的各种变化都可以归纳成受到三类因素的综合影响。

（1）长期趋势波动：它包括长期趋势和无固定周期的循环波动。

（2）季节性变化：它包括所有具有稳定周期的循环波动。

（3）随机波动：除了长期趋势波动和季节性变化之外，其他因素的综合影响归为随机波动。

这三大因素的综合影响会导致序列呈现出各种变化情况。而我们进行确定性时序分析的目的无外乎以下两种：一是克服其他因素的影响，单纯测出其中一个确定性因素（长期趋势波动或季节效应）对序列的影响；二是推断出各种确定性因素彼此之间的相互作用关系及它们对序列的综合影响。

8.1 趋势分析

有些时间序列具有非常显著的趋势，我们分析的目的就是要找到序列中的这种趋势，并利用这种趋势对序列的发展作出合理的预测。

8.1.1 趋势拟合法

趋势拟合法就是把时间作为自变量，相应的序列观察值作为因变量，建立序列值随时间变化的回归模型的方法。根据序列所表现出的线性或非线性特征，我们的拟合

方法又可以具体分为线性拟合和曲线拟合。

1．线性拟合

如果长期趋势呈现出线性特征，那么我们可以用线性模型来拟合它，模型可以具体写为：

$$\begin{cases} T_t = \mathrm{a} + \mathrm{b}t + I_t \\ E(I_t) = 0, \mathrm{Var}(I_t) = \mathrm{c} \end{cases}$$

式中，$\{I_t\}$ 为随机波动；$T_t = \mathrm{a} + \mathrm{b}t$ 就是消除随机波动影响之后该序列的长期趋势。

2．曲线拟合

如果长期趋势呈现出非线性特征，那么我们可以用曲线模型来拟合它。

对曲线模型进行参数估计时，指导思想是：能转换成线性模型的都转换成线性模型，用线性最小二乘法进行参数估计；实在不能转换成线性模型的，就用迭代法进行参数估计。

常用的曲线模型和对应的参数估计方法如表 8-1 所示。

表 8-1　常见曲线模型及参数估计方法

模　　型	变　　换	参数估计方法
二次型：$T_t=\mathrm{a}+\mathrm{b}t+\mathrm{c}t^2$	令 $t_2=t^2$，原模型变换为 $T_t=\mathrm{a}+\mathrm{b}t+\mathrm{c}t_2$	线性最小二乘法
指数型：$T_t=\mathrm{a}\mathrm{b}^t$	对原模型求对数，再令 $T_t'=\ln T_t$,a′=lna,b′=lnb，原模型变换为 $T_t'=\mathrm{a}'+\mathrm{b}'t$	用线性最小二乘法求出 a′，b′，再做变换 $\mathrm{a}=\mathrm{e}^{\mathrm{a}'}$，$\mathrm{b}=\mathrm{e}^{\mathrm{b}'}$
修正指数型：$T_t=\mathrm{a}+\mathrm{b}\mathrm{c}^t$	不能转换成线性模型	迭代法
Gompertz 型：$T_t=\mathrm{e}^{\mathrm{a}+\mathrm{b}\mathrm{c}^t}$	不能转换成线性模型	迭代法
Logistic 型：$T_t=\dfrac{1}{\mathrm{a}+\mathrm{b}\mathrm{c}^t}$	不能转换成线性模型	迭代法

8.1.2　平滑法

平滑法是进行趋势分析和预测时常用的一种方法。它是利用修匀技术，削弱短期随机波动对序列的影响，使序列平滑化，从而显示出变化的规律。它具有调节灵活、

计算简便的特征，广泛被应用于计量经济、人口研究等诸多领域。根据所用的平滑技术的不同，平滑法又可以具体分为移动平均法和指数平滑法。

1．移动平均法

移动平均法的基本思想是：对于一个时间序列$\{X_t\}$，我们可以假定在一个比较短的时间间隔里，序列的取值是比较稳定的，它们之间的差异主要是由随机波动造成的。根据这种假定，可以用一定时间间隔内的平均值作为某一期的估计值。它又可以分为两类：

（1）n 期中心移动平均

$$\tilde{X}_t=\begin{cases}\dfrac{1}{n}\left(X_{t-\frac{n-1}{2}}+X_{t-\frac{n-1}{2}+1}+\cdots+X_t+\cdots+X_{t+\frac{n-1}{2}-1}+\dfrac{1}{2}X_{t+\frac{n-1}{2}}\right), & n\text{为奇数}\\ \dfrac{1}{n}\left(\dfrac{1}{2}X_{t-\frac{n}{2}}+X_{t-\frac{n}{2}+1}+\cdots+X_t+\cdots+X_{t+\frac{n}{2}-1}+\dfrac{1}{2}X_{t+\frac{n}{2}}\right), & n\text{为偶数}\end{cases}$$

（2）n 期移动平均

$$\tilde{X}_t=\frac{1}{n}(X_t+X_{t-1}+\cdots+X_{t-n+1})$$

移动平均的期数对原序列的修匀效果影响很大，要确定移动平均的期数，一般会从以下三个方面加以考虑。

① 事件的发展有无周期性。如果事件的发展具有一定的周期性，一般以周期长度作为移动平均的间隔长度。比如研究每月的平均气温变化趋势，就应该做 12 期移动平均，通过周期平滑消除季节效应的影响。

② 我们对趋势平滑性的要求。一般移动平均的期数越多，修匀曲线越平滑，表现出的长期趋势就越清晰。

③ 我们对趋势反映近期变化敏感程度的要求。用移动平均方法确定事件的发展趋势都具有一定的滞后性，移动平均的期数越多，滞后性越大，移动平均的期数越少，所得的趋势图对近期变化的反应就越敏感。

综合如上方面的考虑，如果想得到长期趋势，就应该做期数较大的移动平均，如

果想密切关注序列的短期趋势，就应该做期数较小的移动平均。

在预测领域，n 期移动平均还是一种常用的预测方法。假定最后一期的观察值为 X_T，那么使用 n 期移动平均方法，向前 l 期的预测值为：

$$\tilde{X}_t = \frac{1}{n}(X_t + X_{t-1} + \cdots + X_{t-n+1})$$

式中

$$X'_{T+l-i} = \begin{cases} \tilde{X}_{T+l-i}, l > i \\ X_{T+l-i}, l \leqslant i \end{cases}, i = 1, 2, \cdots, n$$

2．指数平滑法

移动平均法实际上就是用一个简单的加权平均数作为某一期趋势的估计值。以 n 期移动平均为例，$\tilde{X}_t = \frac{1}{n}(X_t + X_{t-1} + \cdots + X_{t-n+1})$，相当于用近 n 期的加权平均数作为最后一期趋势的估计值，它们的权重都取为 $\frac{1}{n}$，实际上也就是假定无论时间远近，这 n 期的观察值 $X_t, X_{t-1}, \cdots, X_{t-n+1}$ 对最后一期的影响都是一样的。但在实际生活中，我们会发现对大多数随机事件而言，一般都是近期的结果对现在的影响会大些，远期的结果对现在的影响会小些。为了更好地反映这种影响作用，我们将考虑到时间间隔对事件发展的影响，各期权重随时间间隔的增大而呈指数衰减，这就是指数平滑法的基本思想。不考虑季节（周期）因素的影响，常用的进行趋势拟合的简单指数平滑公式如下：

$$\tilde{X}_t = \alpha X_t + \alpha(1-\alpha)X_{t-1} + \alpha(1-\alpha)^2 X_{t-2} + \cdots$$

式中，α 为平滑系数，它满足 $0 < \alpha < 1$。

因为：

$$\tilde{X}_{t-1} = \alpha X_{t-1} + \alpha(1-\alpha)X_{t-2} + \alpha(1-\alpha)^2 X_{t-3} + \cdots$$

所以 $\tilde{X}_t$ 又等价于：

$$\tilde{X}_t = \alpha X_t + (1-\alpha)\tilde{X}_{t-1}$$

简单指数平滑面临一个确定 $\tilde{X}_0$ 初始值的问题。我们有许多方法可以确定 $\tilde{X}_0$ 的初始值，最简单的方法是指定 $\tilde{X}_0 = X_1$。

平滑系数 α 的值由研究人员根据经验给出。一般对于变化缓慢的序列，常取较小的 α 值；相反对于变化迅速的序列，常取较大的 α 值。经验表明 α 的值介于 0.05～0.3 之间，修匀效果比较好。

指数平滑法也是一种常用的预测方法，$\tilde{X}_T$ 常常作为 1 期预测值：

$$\hat{X}_{T+1} = \tilde{X}_T = \alpha X_T + \alpha(1-\alpha)X_{T-1} + \alpha(1-\alpha)^2 X_{T-2} + \cdots = \alpha X_T + (1-\alpha)\hat{X}_{T-1}$$

指数平滑 2 期预测值为：

$$\hat{X}_{T+2} = \alpha\hat{X}_{T+1} + \alpha(1-\alpha)X_T + \alpha(1-\alpha)^2 X_{T-1} + \cdots = \alpha\hat{X}_{T+1} + (1-\alpha)\hat{X}_{T+1} = \hat{X}_{T+1}$$

容易验证，指数平滑 l 期预测值都具有如下关系：

$$\hat{X}_{T+l} = \hat{X}_{T+1}, l \geqslant 2$$

8.2 季节效应分析

在日常生活中，我们可以见到许多有季节效应的时间序列，比如气温、每个月的商品零售额、某自然景点每季度的旅行人数等，它们都会呈现出明显的季节变动规律。我们还可以把“季节”广义化，凡是呈现出固定的周期性变化的事件，我们都称它具有“季节”效应。现在“季节”效应已经变成周期效应的代名词，而“季”也变成周期内每一期的代名词。

8.2.1 季节性时间序列模型

所谓季节性时间序列，是指具有某种周期性变化规律的随机序列，并且这种周期性的变化规律往往是由于季节变化引起的。

如果一个随机序列经过 S 个时间间隔后观测数据呈现相似性，比如同处于波峰或波谷，则我们称该序列具有以 S 为周期的周期特征，并称其为季节性时间序列，S 为季节长度。

季节性时间序列表现出同期相关性，也就是说时间相隔为 S 的两个时间点上的随机变量有较强的相关性。比如，对于月度数据 S=12，则 X_t 与 X_{t-12} 相关性较强。我们可以利用这种同期相关性在 X_t 与 X_{t-12} 之间进行拟合。

简单季节模型通过简单的趋势差分、季节差分之后序列即可转化为平稳，它的模型结构通常表示如下：

$$\Phi\left(L^S\right)\left(1-L^S\right)^D X_t=\Theta\left(L^S\right)\varepsilon_t$$

其中 $\{\varepsilon_t\}$ 为白噪声序列。

其中的两个函数可以定义为：

$$\Phi\left(L^S\right)=1-\Phi_1 L^S-\Phi_2 L^{2S}-\cdots\Phi_p L^{pS},\ \text{SAR 算子}$$

$$\Theta\left(L^S\right)=1-\Theta_1 L^S-\Theta_2 L^{2S}-\cdots\Theta_p L^{qS}.\ \text{SMA 算子}$$

该模型称为简单季节模型或季节性自回归求和移动平均模型，简记为SARIMA $\left(p,D,q\right)_S$ 模型。

2008—2017 年 GDP 的季度数据存在很强的季节效应，如图 8-1 所示。

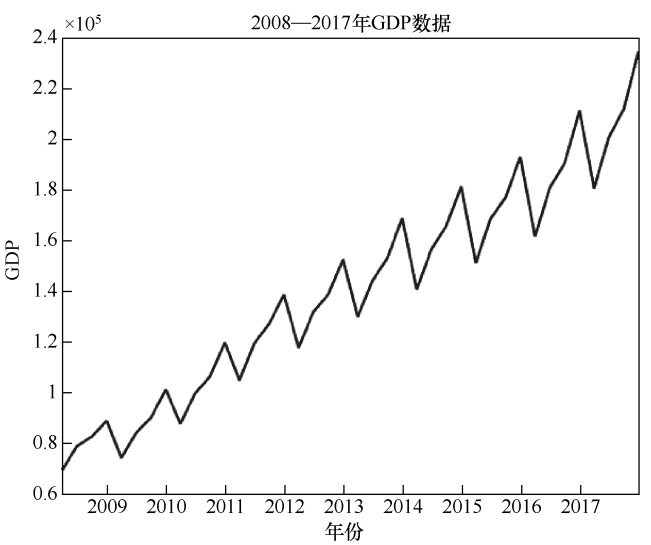


图 8-1　2008—2017 年 GDP 季度数据

8.2.2 季节指数

所谓季节指数就是用简单平均法计算的周期内各时期季节性影响的相对数，考虑季节模型：

$$X_{ij} = \bar{X}S_j + I_{ij}$$

季节指数的计算分为三步。

第一步：计算周期内各期平均数，得到长期以来该时期的平均水平。假定序列的数据结构为 m 期为一周期，共有 n 个周期。则：

$$\bar{X}_k = \frac{\sum_{i=1}^{n} X_{ik}}{n}, k = 1,2,\cdots,m$$

第二步：计算总平均数。

$$\bar{X} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{m} X_{ik}}{nm}$$

第三步：用时期平均数除以总平均数就可以得到各时期的季节指数 $S_k(k=1,2,\cdots,m)$，即：

$$S_k = \frac{\bar{X}_k}{\bar{X}}, k = 1,2,\cdots,m$$

这就是季节指数的构造方法。季节指数反映了该季度与总平均值之间的一种比较稳定的关系。如果这个比值大于 1，就说明该季度的值常常会高于总平均值；如果这个比值小于 1，就说明该季度的值常常低于总平均值；如果序列的季节指数都近似等于 1，那就说明该序列没有明显的季节效应。

8.2.3 季节效应的差分方法

对于有趋势性和季节成分的时间序列，差分运算是一种常见的剔除趋势性和季节成分的方法。差分运算即简单地取两个观测值的差值。设时间序列为 X_t，则滞后 1 期差分（又称为 1 步差分）是取时间序列中每两个相邻的观测值的差值，即 $\Delta X_t =$

$X_t - X_{t-1}$。滞后 k 期差分（又称为 k 步差分）是指每相隔 k 个时间点的观测值之间的差值，即当前观测值和退后 k 期的观测值之间的差值，记为 $\Delta_k X_t = X_t - X_{t-k}$。例如，一个日时间序列 X_t，滞后 7 期（或者称为 7 步）差分意味着每天的观测值 X_t 减去上一周同一天的观测值 X_{t-7}，即 $\Delta_7 X_t = X_t - X_{t-7}$。

剔除趋势和季节性成分的一种方法是进行差分运算。对原始时间序列经过适当的差分运算，得到的差分序列一般不再有趋势性或者季节成分。原始时间序列进行一阶差分之后得到的序列就是一阶差分序列，它衡量时间序列从一个时间点到相邻时间点的变化。

一般情况下，进行一阶差分就可以剔除线性趋势。通过差分来剔除趋势成分不需要假设该趋势是全局性的，因为整个时间范围内的差分运算都是一致的。如果通过时间序列来拟合趋势成分的话，那么需要假定该趋势成分是全局性的趋势，即该趋势也适用于未来需要预测的时期，这一点是差分运算优于用回归来拟合趋势的地方。

对于二次项趋势和指数趋势，一阶差分不能够完全剔除趋势，它需要在一阶差分序列的基础上继续进行一阶差分运算，得到的序列称为二阶差分序列。如果二阶差分序列还是含有趋势成分的话，则继续进行差分运算，依此类推。

对于一个时间序列，如果它含有季节趋势，季节的个数设为 M。一般情况下，对该时间序列进行滞后 M 期季节差分即可以剔除季节趋势。例如，如果一个日数据含有周季节趋势，这里季节的个数为 7，对该序列进行 7 阶滞后季节差分则可以剔除周季节趋势。

如果一个时间序列既有趋势性，又有季节性成分，可以进行两次差分运算来分别剔除趋势性和季节性成分。

8.3 模型的应用实例

8.3.1 线性趋势拟合模型

可以使用线性趋势模型拟合中国总人口，并分析其趋势，我们使用的数据是 1949—2016 年中国的总人口。自变量为年份，因变量为人口数（万人），使用 MATLAB 函

数‘polyfit’或者‘Curve Fitting’均可进行线性拟合，模型的实现脚本如下。

（1）读取数据

```
clc, clear all, close all
data=xlsread('populationdata','Sheet1','A2:B69');
population_num = data(:,2);
Dates = data(:,1);
N = length(population_num);
```

（2）数据可视化

```
figure(1)
plot(Dates,population_num,'LineWidth',1.5);
title('1949年至2016年中国总人口')
ylabel('人口数(万人)')
xlabel('年份')
```

如图 8-2 所示，可以看到 1949—2016 年中国总人口数据有很明显的向上趋势性。

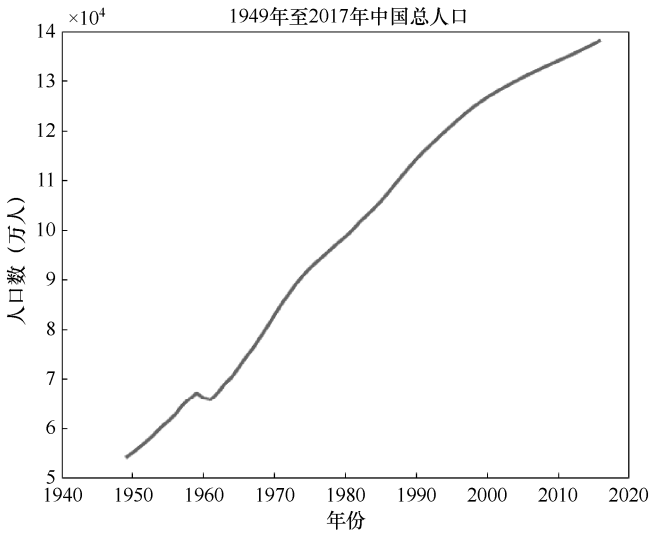


图 8-2　中国总人口数据

（3）拟合及画图

```
P = polyfit(Dates,population_num,1);
popu_fore = P(1)*Dates+P(2);
figure(2)
plot(Dates,population_num,'k--','LineWidth',1.5);
```

```
hold on
plot(Dates,popu_fore,'b','LineWidth',1.5);
title('1949年至2016年中国总人口')
ylabel('人口数（万人）')
xlabel('年份')
xlim([1949,2016])
legend('实际人口数','拟合曲线')
```

拟合的结果为：

```
Linear model Poly1:
    f(x) = p1*x + p2
Coefficients (with 95% confidence bounds):
      p1 =  1366  (1327, 1405)
      p2 =  -2.609e+06  (-2.686e+06, -2.531e+06)

Goodness of fit:
  SSE: 6.63e+08
  R-square: 0.9866
  Adjusted R-square: 0.9864
  RMSE: 3169
```

因此拟合方程为 $f(t)=1366t-2\,609\,000$，检验结果表示该拟合成立，图像如图 8-3 所示。

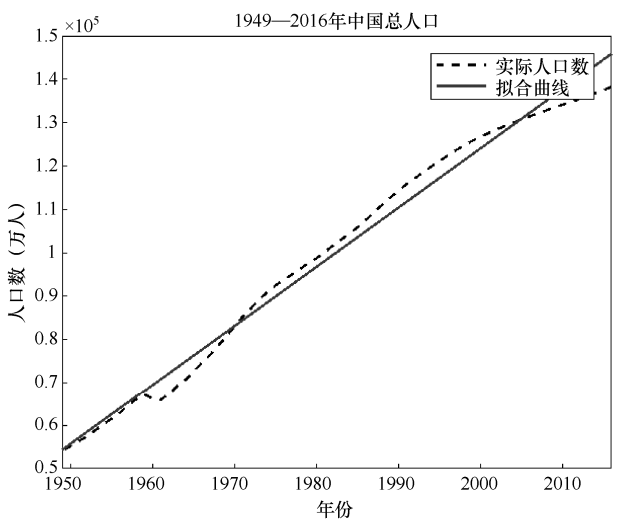


图 8-3　中国总人口曲线拟合图

8.3.2 季节性建模

如图 8-1 所示，中国的季度 GDP 数据有很强的季节效应，因此对该时间序列的分析需要建立季节性模型，我们使用的数据是 1992—2017 年的季度 GDP 数据，模型的实现脚本如下。

（1）读取数据

```
clc, clear all, close all
[GDP,date]=xlsread('data','Sheet1','A2:B105');
N = length(GDP);
a = zeros(N,1);
for i = 1:N
   a(i) = datenum(cell2mat(date(i)));
end
a = datestr(a);
a = datevec(a);
Dates = datetime(a);
```

（2）数据可视化

```
figure(1)
plot(Dates,GDP,'LineWidth',1.5);
title('1992 年至 2017 年中国 GDP')
ylabel('GDP')
xlabel('日期')
```

如图 8-4 所示，该数据有明显的趋势性和季节性。

（3）差分

```
diffGDP = zeros(N-4,1);
for i = 1:N-4
   diffGDP(i) = GDP(i+4)-GDP(i);
end
diff_2 = diff(diffGDP);
figure(2)
plot(diff_2,'LineWidth',1.5);
title('1992 年至 2017 年中国 GDP 的二阶差分数据')
[hp,hpValue,stat,cValue,reg] = pptest(diff_2,'model','TS')
```

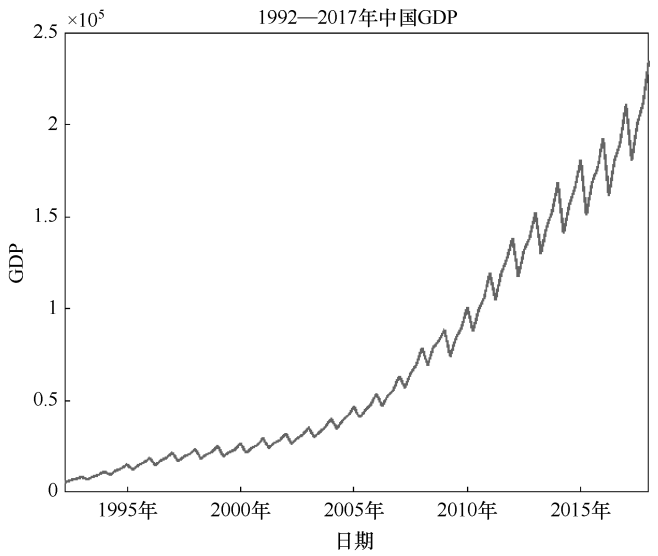


图 8-4　1992—2017 年中国季度 GDP

我们通过两次差分来去除其趋势季节性，差分后的数据如图 8-5 所示。

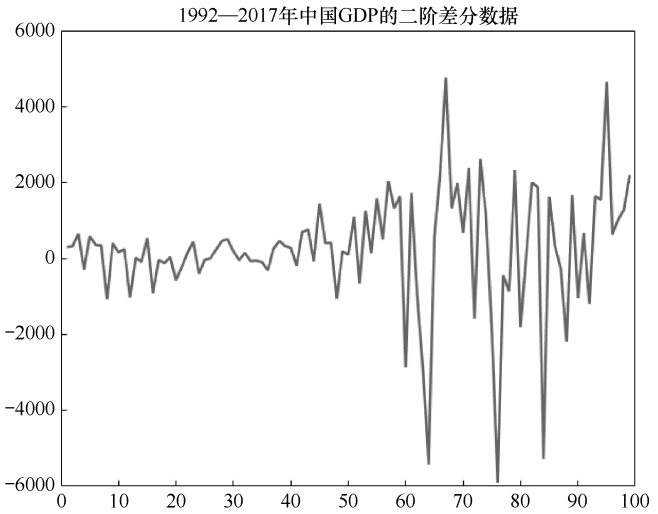


图 8-5　中国季度 GDP 数据二阶差分

对该数据进行 pp 检验，可得

```
    hp =
```

```
  logical

   1


hpValue =

   1.0000e-03
```

pp 检验的结果说明该模型二阶差分后的数据平稳。

（4）AIC 定阶

```
maxLags = 4;
AICSet = zeros(maxLags, maxLags);

for i = 1:maxLags
for j = 1:maxLags
       mdl = arima('ARLags',[1:i],'MALags',[1:j]);
       [EstMdl, EstParamCov, logL, info] = estimate(mdl, diff_2,
       'display', 'off');
       AICSet(i, j) = aicbic(logL, length(info.X));

end
end

% 画热度图来表示 AIC 数值的分布
figure(3)
heatmap(AICSet/1000);
xlabel('MA Lags')
ylabel('AR Lags')
title('AIC 准则图')

[OptimalARLags, OptimalMALags] = find(AICSet==min(min(AICSet)));
```

对差分后的该数据进行 AIC 定阶分布如图 8-6 所示，可以看到最适合的模型是 ARMA（2,3）。

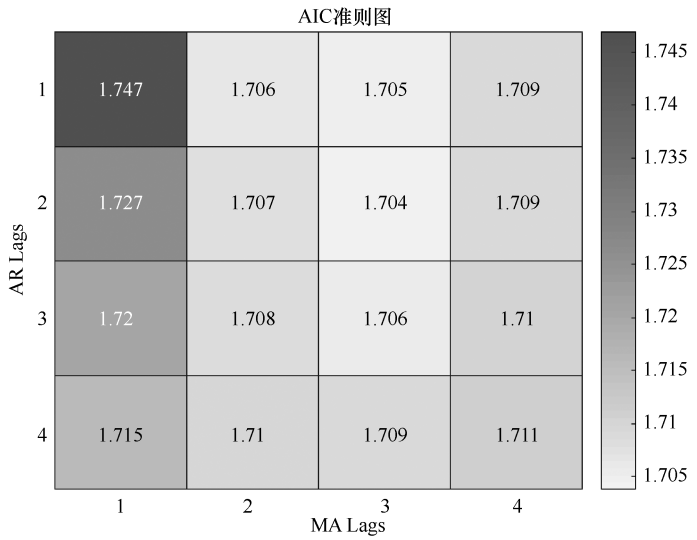


图 8-6　AIC 准则定阶分布图

（5）建立模型

```
mdl = arima(OptimalARLags,0, OptimalMALags);
fit = estimate(mdl, diff_2);
```

建立模型结果如下：

```
ARIMA(2,0,3) Model:
   --------------------
   Conditional Probability Distribution: Gaussian

                            Standard          t
    Parameter       Value       Error     Statistic
   -----------   -----------   ------------   -----------
    Constant        53.3867       77.962      0.684779
       AR{1}      -0.170179     0.141462        -1.203
       AR{2}       0.804558     0.134172       5.99645
       MA{1}       0.765293     0.107529       7.11706
       MA{2}       -0.88589     0.136902        -6.471
       MA{3}      -0.879403     0.108577      -8.09934
    Variance    2.38797e+06       522558       4.56978
```

因此模型形式为：

$$\Phi(L^2)(1-L^4)^2 X_t = \Theta(L^3)\varepsilon_t$$

（6）直接建立季节模型的 AIC 定阶

此外还可以在 MATLAB 中直接利用 arima 函数中的‘Seasonality’来设置季节性模型。我们直接对 GDP 数据建模，使用一阶差分去除数据的趋势性，但是如果使用‘Seasonality’设定季节参数，需要使用 AIC 准则判断其最优阶数。

```
maxLags = 4;
AICSet = zeros(maxLags, maxLags);

for i = 1:maxLags
for j = 1:maxLags
      mdl = arima('D',1,'Seasonality',4,'SARLags',i,'SMALags',j);
      [EstMdl, EstParamCov, logL, info] = estimate(mdl, GDP, 'dis-
      play', 'off');
      AICSet(i, j) = aicbic(logL, length(info.X));

end
end

% 画热度图来表示 AIC 数值的分布
figure(4)
heatmap(AICSet/1000);
xlabel('MA Lags')
ylabel('AR Lags')
title('Akaike information criteria')

[OptimalARLags, OptimalMALags] = find(AICSet==min(min(AICSet)));
title(['Optimal AR and MA Lags are (' num2str(OptimalARLags) ','
num2str(OptimalMALags) ')'])
```

如图 8-7 所示，模型的最优的阶数为（4，4）。

（7）季节性模型建模

```
mdl = arima('D',1,'Seasonality',4,'SARLags',OptimalARLags,'SMAL-
ags',OptimalMALags);
fit = estimate(mdl, GDP);
```

```
GDP = GDP - mean(GDP);
disp(['检验残差是否存在相关性']);
[hLBQ, pLBQ] = lbqtest(GDP, 'Lags', 1:4, 'alpha', 0.05)
```

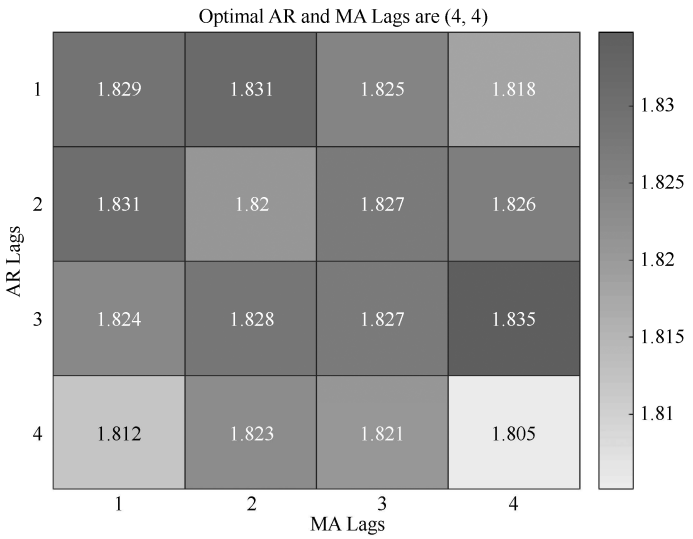


图 8-7　季节性模型 AIC 准则定阶分布图

直接建立季节性模型有如下结果：

```
ARIMA(0,1,0) Model Seasonally Integrated with Seasonal AR(4) and
MA(4):
-----------------------------------------------------------------
Conditional Probability Distribution: Gaussian

                          Standard          t
   Parameter      Value         Error      Statistic
-----------   -----------   ------------   -----------
Constant        7.9549       38.0271       0.20919
  SAR{4}             1     0.0867796       11.5234
  SMA{4}      -0.871171      0.176264      -4.94242
Variance    2.27243e+06        306534        7.4133
```

用 LBQ 检验判断残差相关性结果如下，可知残差没有相关性。

```
hLBQ =
```

```
    1×4 logical array

     1   1   1   1

pLBQ =

     0     0     0     0
```

（8）季节性模型预测

```
[Yf,YMSE] = forecast(fit,10,'Y0',GDP);
upper = Yf + 1.96*sqrt(YMSE);
lower = Yf - 1.96*sqrt(YMSE);

figure(5)
plot(GDP,'b','LineWidth',2);
hold on
h1 = plot(N+1:N+10,Yf,'r','LineWidth',2);
title('GDP 模型及预测')
legend('实际值','预测值')
hold off
```

我们利用该模型对未来数据进行预测，结果如图 8-8 所示，实线为实际值，虚线为预测值，符合未来发展趋势。

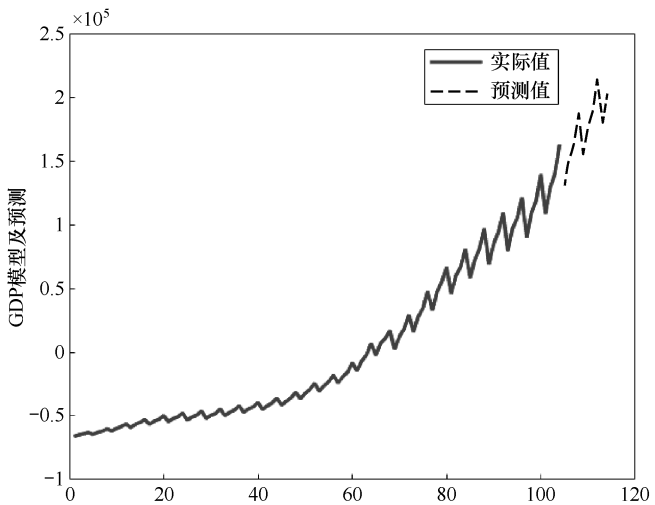


图 8-8　季节性模型预测

8.4 小结

在现实生活中产生的非平稳序列可能显示出非常明显的规律性，比如有显著的趋势或者有固定的变化周期，本章主要介绍趋势及季节性时间序列建模的过程。对于趋势数据，我们可以用时间作为自变量，相应的观察序列作为因变量，建立序列值随时间变化的回归模型，也可以对趋势数据进行平滑化，削弱短期随机波动对序列的影响。对于季节效应的时间序列数据在生活中随处可见，凡是呈现出固定的周期性变化的事件，都称它具有“季节”效应。我们可以利用季节指数，用简单平均法计算周期内各时期季节性影响的相对数。

此外差分方法是处理季节性和趋势性的一种很好的方法，一般情况下，进行一阶差分就可以剔除线性趋势。对于一个时间序列，如果它含有季节趋势，季节的个数设为 M。一般情况下，对该时间序列进行滞后 M 期季节差分即可以剔除季节趋势。在 8.3 节我们给出了线性趋势拟合和季节性模型的实例。在 MATLAB 中，建立季节性模型，既可以通过差分消除季节性后再建模，也可以利用函数 arima 中的参数直接设定季节性时间序列模型。

参考文献

[1] Tseng F M, Tzeng G H. A fuzzy seasonal ARIMA model for forecasting[J]. Fuzzy Sets & Systems, 2002, 126(3):367-376.

[2] Hillmer S C, Tiao G C. An ARIMA-Model-Based Approach to Seasonal Adjustment[J]. Publications of the American Statistical Association, 1982, 77(377):63-70.

[3] Jonathan D.Cryer, Kung-Sik Chan. Time Series Analysis With Applications in R (Second Edition): New York, Springer, 2008.

[4] 王黎明，王连，杨楠. 应用时间序列分析. 上海：复旦大学出版社，2009.

[5] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[6] 孙祝岭. 时间序列与多元统计分析. 上海：上海交通大学出版社，2016.

[7] 胡永宏，王振龙. 应用时间序列分析. 北京：科学出版社，2007.

9 条件异方差模型

1982 年，Engle 在分析英国通货膨胀率序列时，发现经典的 ARIMA 模型始终无法取得理想的拟合效果。经过对残差序列仔细的研究，他发现问题出在残差序列具有异方差性。本章主要介绍条件异方差模型。

9.1 时间序列的异方差性

9.1.1 异方差性

使用 ARIMA 模型拟合非平稳序列时，对残差序列有一个重要假定——残差序列 $\{\varepsilon_t\}$ 为零均值白噪声序列。换言之，残差序列要满足如下三个假定条件。

① 零均值

$$E(\varepsilon_t)=0$$

② 纯随机

$$\mathrm{Cov}(\varepsilon_t,\varepsilon_{t-i})=0,\forall i\geqslant 1$$

③ 方差齐性

$$\mathrm{Var}(\varepsilon_t)=\sigma_t^2$$

如果方差齐性假定不成立，即随机误差序列的方差不再是常数了，它会随着时间的变化而变化，可以表示为时间的某个函数：

$$\mathrm{Var}(\varepsilon_t)=h(t)$$

这种情况被称作异方差。

在残差序列的这三个假定中，零均值假定最容易实现，只要对序列进行中心化处理就可以实现，所以这个假定通常无须检验。

纯随机假定一直是我们重点监控的对象，如果这个假定不满足，就说明残差序列中还蕴含着没有提取的自相关信息。

只有第三个假定——方差齐性假定，在本章之前我们没有进行任何检验。在缺省检验的情况下就默认残差序列一定满足这个条件。但实际上，这个假定条件并不总是满足，忽视异方差的存在会导致残差的方差被严重低估，继而参数显著性检验容易犯纳伪错误，这使得参数的显著性检验失去意义，最终导致模型的拟合精度受影响。所以为了提高模型拟合的精度，我们需要对残差序列进行方差齐性检验，并且对异方差序列进行深入分析。

9.1.2 异方差的残差图

当残差序列$\{\varepsilon_i\}$方差齐性时，它应该在零值附近随机波动，不带任何趋势，否则就显示出异方差的性质了，如图 9-1 所示。

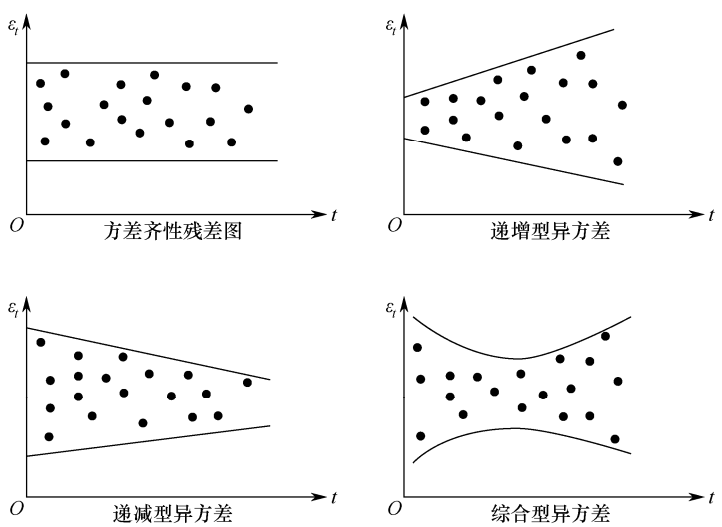


图 9-1　残差序列示意图

由于残差序列的方差实际上就是它平方的期望，即：

$$\mathrm{Var}(\varepsilon_t) = E(\varepsilon_t^2)$$

所以残差序列是否方差齐性主要是考察 ε_t^2 的性质。我们可以借助残差平方图——ε_t^2 关于 t 变化的二维坐标图，对残差序列的方差齐性进行直观诊断。

和残差图的判断原则一样，假设方差齐性满足的话，有：

$$E(\varepsilon_t^2) = \sigma_t^2$$

这意味着 ε_t^2 应该在某个常数值 σ_t^2 附近随机波动，它不应该具有任何明显的趋势，否则就呈现出异方差性。

当残差序列存在异方差时，我们需要对它进行进一步的处理，处理思路有两种：

① 假如已知异方差函数具体形式，进行方差齐性变化。

② 假如不知异方差函数的具体形式，拟合条件异方差模型。

9.1.3 方差齐性变换

假设序列显示出显著的异方差性，且方差 σ_t^2 与均值 μ_t 之间具有某种函数关系：

$$\sigma_t^2 = h(\mu_t)$$

式中，$h(u_t)$ 是某个已知函数。

在这种情况下，我们的处理思路是尝试寻找一个转换函数 $g(x)$，使得经转换后的变量 $g(x_t)$ 满足方差齐性：

$$\mathrm{Var}\left[g(x_t)\right] = \sigma^2$$

恩格尔和克拉格（Kraft,D.,1983）在分析宏观数据时，发现这样一种现象：时间序列模型中的扰动方差稳定性比通常假设的要差。恩格尔的结论说明在分析通货膨胀模型时，大的及小的预测误差会大量出现，表面存在一种异方差，其中预测误差的方差取决于后续扰动项的大小。

那些从事股票价格、通货膨胀率、外汇汇率等金融时间序列预测工作的研究者，曾发现他们对这些变量的预测能力随时期的不同有相当大的变化。预测的误差在某一

时期里相对较小，而在某一时期里则相对较大，然后，在另一时期又是较小的。这种变异很可能由于金融市场的波动性易受谣言、政局变动、政府货币与财政政策变化等影响，从而说明预测误差的方差中有某种相关性。

为了刻画这种相关性，恩格尔提出自回归条件异方差（ARCH）模型。ARCH 模型的主要思想是时刻 t 的 ε_t 的方差 (σ_t^2) 依赖于时刻 $t-1$ 的残差平方的大小，即依赖于 ε_{t-1}^2。

9.2 异方差性检验

考虑残差序列 $\{\varepsilon_t\}$ 的异方差性，异方差检验的实质是进行异方差相关性检验。最常用的两个异方差检验方法是 Portmantea Q 检验和 LM 检验。

9.2.1 Portmantea Q 检验

1983 年 Mcleod 和 Li 在 LB 统计量的基础上提出了 Portmanteau Q 统计量，用于检验残差平方序列的自相关性。

该检验的假设条件为：

$$\mathrm{H}_0:\text{残差平方序列纯随机} \leftrightarrow \mathrm{H}_1:\text{残差平方序列具有自相关性}$$

用自相关性系数表示，该假设条件等价于：

$$\mathrm{H}_0:\rho_1=\rho_2=\cdots=\rho_q=0 \leftrightarrow \mathrm{H}_1:\rho_1,\rho_2,\cdots,\rho_q \text{ 不全为零}$$

检验统计量为：

$$`Q(q)=n(n+2)\sum_{i=1}^{q}\frac{\rho_i^2}{n-i}$$

式中，n 为观察序列长度；ρ_i 为残差序列延迟 i 的自相关系数，记：

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\varepsilon_t^2}{n}$$

$$\rho_i=\sqrt{\frac{\sum_{t-i=1}^{n}(\varepsilon_t^2-\hat{\sigma}^2)(\varepsilon_{t-i}^2-\hat{\sigma}^2)}{\sum_{t=1}^{n}(\varepsilon_t^2-\hat{\sigma}^2)^2}}$$

Portmanteau Q 统计量近似服从自由度为 $q-1$ 的 χ^2 分布：

$$Q(q):\chi^2(q-1)$$

检验结果如表 9-1 所示。

表 9-1　Portmanteau Q 统计量检验结果

显著性水平	临界值	Portmanteau Q 统计量	*P* 值	检验结果
α	$\chi^2_{1-\alpha}(q-1)$	$Q(q)\geqslant\chi^2_{1-\alpha}(q-1)$	$\leqslant\alpha$	拒绝 H_0
		$Q(q)\geqslant\chi^2_{1-\alpha}(q-1)$	$>\alpha$	接受 H_0

9.2.2　拉格朗日乘子检验

1982 年，Engle 为了确定 ARCH(q)模型的阶，提出了拉格朗日乘子检验（Lagrange Multiplier Test），简记为 LM 检验。

假设条件为：

H_0 :残差平方序列纯随机 $\leftrightarrow$ H_1 :残差平方序列具有自相关性

等价于：

$$H_0:\rho_1=\rho_2=\cdots=\rho_q=0\leftrightarrow H_1:\rho_1,\rho_2,\cdots,\rho_q\text{ 不全为零}$$

可以证明若 H_0 为真，则：

$$LM(q)=W^TW$$

式中：

$$W=\left(\frac{\rho_1^2}{\sigma^2},\frac{\rho_2^2}{\sigma^2},\cdots,\frac{\rho_q^2}{\sigma^2}\right)$$

$$\rho_t = \sqrt{\frac{\sum_{t=i+1}^{n}(\varepsilon_t^2 - \hat{\sigma}^2)(\varepsilon_{t-1}^2 - \hat{\sigma}^2)}{\sum_{i=1}^{n}(\varepsilon_t^2 - \hat{\sigma}^2)^2}}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\varepsilon_t^2}{n}$$

LM（q）统计量近似服从自由度为 $q-1$ 的 χ^2 分布。

9.3 自回归条件异方差模型

9.3.1 ARCH 模型的原理

方差齐性变换为异方差序列的精确拟合提供了一个很好的解决方法，但这个方法更多的只具有理论上的意义，在实践中的可操作性并不强。因为要使用方差齐性变化必须要事先知道异方差函数的形式，而这通常是不可能的。

实践中，我们只能根据残差图及残差平方图所显示出来的特点，使用一些常用的函数形式估计异方差函数。这体现在进行金融时间序列分析时，由于金融序列的标准差与水平值之间通常具有某种正相关关系，所以异方差函数经常被假定为：

$$h(\varepsilon_t) = \varepsilon_t^2$$

进而导致对数变换在进行金融时间序列分析时被普遍采用。但大量的实践证明这种假定太简单化了，对数变换通常只能使绝大多数金融时间序列的异方差程度得到改善，但无法真正实现方差齐性。为了更精确地估计异方差函数，Engle 于 1982 年提出了条件异方差模型。

ARCH 模型的全称是自回归条件异方差模型（Autoregressive Conditional Heteroskedastic），它的构造原理如下：

假设在历史数据已知的情况下，零均值、纯随机残差序列具有异方差性：

$$\mathrm{Var}(\varepsilon_t) = h_t$$

在正态分布的假定下，有：

$$\varepsilon_t \Big/ \sqrt{h_t} \sim \mathrm{N}(0,1)$$

异方差等价于残差平方的均值：

$$E(\varepsilon_t^2) = h_t$$

使用残差平方序列的自相关系数，可以考察异方差函数的自相关性：

$$\rho_t = \sqrt{\frac{\mathrm{Cov}(\varepsilon_t^2, \varepsilon_{t-i}^2)}{\mathrm{Var}(\varepsilon_t^2)}}$$

若 $\rho_{\mathrm{t}} \neq 0, t = 1,2,\cdots$，误差平方序列的自相关系数不恒为零，说明异方差函数存在自相关性，这使得我们可以通过构造残差平方序列的自回归模型来拟合异方差函数。

为了说得更具体，让我们回到 k–变量回归模型：

$$y_t = a_0 + a_1 x_{1t} + \cdots + a_k x_{kt} + \varepsilon_t$$

并假设在时刻 $t-1$ 所有信息已知的条件下，扰动项 ε_t 的分布是：

$$\varepsilon_t : \mathrm{N}(0, (\alpha_0 + \alpha_1 \varepsilon_{t-1}^2))$$

也就是，ε_t 遵循以 0 为均值，$(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)$ 为方差的正态分布。

由于 ε_t 的方差依赖于前期的平方扰动项，我们称它为 ARCH（1）过程：

$$\mathrm{Var}(\varepsilon_t) = \sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$$

加以推广可得一个 ARCH（p）过程，可以写为：

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2$$

如果扰动项方差中没有自相关，就会有：

$$\mathrm{H}_0 : \mathrm{Var}(\varepsilon_t) = \sigma^2 = \alpha_0$$

这时：

$$\alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$$

从而得到误差的同方差性情形。

恩格尔曾表明，可以通过以下的回归去检验上述虚假假设：

$$\hat{\varepsilon}_t^2 = \hat{\alpha}_0 + \hat{\alpha}_1 \hat{\varepsilon}_{t-1}^2 + \hat{\alpha}_2 \hat{\varepsilon}_{t-2}^2 + \cdots + \hat{\alpha}_p \hat{\varepsilon}_{t-p}^2$$

其中，$\hat{\varepsilon}_t$ 表示从原始回归模型估计得到的 OLS 残差。

9.3.2 ARCH 模型的平稳性条件

在 ARCH（1）模型中，观察参数 α 的含义：

当 $\alpha \to 1$ 时，$\mathrm{Var}(\varepsilon_t) \to \infty$；

当 $\alpha \to 0$ 时，退化为传统情形：$\varepsilon_t : \mathrm{N}(0, \omega)$；

ARCH 模型的平稳性条件：$\sum \alpha_i < 1$（这样才得到有限的方差）。

但是若出现 $\rho_k = 0, k = 1, 2, \cdots$ 这说明异方差函数纯随机。此时，历史数据对未来异方差的估计一点作用都没有，这是最难分析的一种情况，目前没有有效的方法提取其中的异方差信息。

9.4 广义自回归条件异方差模型

ARCH 模型的实质是使用误差平方序列的 q 阶移动平均来拟合当期异方差函数值。由于移动平均模型具有自相关系数 q 阶截尾性，所以 ARCH 模型实际上只适用于异方差函数短期自相关过程。

但是在实践中，有些残差序列的异方差函数是具有长期自相关性的，这时如果使用 ARCH 模型拟合异方差函数，将会产生很高的移动平均阶数，这会增加参数估计的难度，并最终影响 ARCH 模型的拟合精度。

我们常常认为 ε_t 的方差依赖于很多时刻之前的变化量（特别是在金融领域，采用日数据或周数据的应用更是如此）。这里的问题在于，我们必须估计很多参数，而这

一点很难精确做到。但是如果我们能够用一个或两个 σ_t^2 的滞后值代替许多 ε_t^2 的滞后值，这就是广义自回归条件异方差模型（Generalized Autoregressive Conditional Heteroscedasticity，GARCH 模型）。在 GARCH 模型中，要考虑两个不同的设定：一个是条件均值，另一个是条件方差。

在标准化的 GARCH（1,1）模型中：

$$y_t = x_t\gamma + \varepsilon_t$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

其中，x_t 是自变量向量；γ 是系数向量。第一个式子中给出的均值方程是一个带有误差项的变量函数。由于 σ_t^2 是以前面信息为基础的一期向前预测方差，所以它被称作条件方差。

该模型给出的条件方差方程是下面三项的函数：

① 常数项（均值）：ω。

② 用均值方程的残差平方的滞后来度量从前期得到的波动性的信息：ε_{t-1}^2（ARCH 项）。

③ 上一期的预测方差：σ_{t-1}^2（GARCH 项）。

GARCH(1,1)模型中的(1,1)是指阶数为 1 的 GARCH 项（括号中的第一项）和阶数为 1 的 ARCH 项（括号中的第二项）。一个普通的 ARCH 模型是 GARCH 模型的一个特例，即在条件方差方程中不存在滞后预测方差 σ_t^2。

9.5 模型的 MATLAB 方法

9.5.1 残差异方差性检验

在MATLAB中检验残差异方差性使用的函数是archtest，该函数的使用方法如下。

h = archtest(res)

h = archtest(res,Name,Value)

[h,pValue] = archtest(___)

[h,pValue,stat,cValue] = archtest(___)

（1）输入变量的含义

- res——用于检验的残差序列，丢失数据用 NaN 代替。

（2）可选择输入变量的含义

- Lags——检验的滞后项数目，可以是正整数或者正整数向量，默认值为 1。
- Alpha——假设检验的显著性水平，默认值为 0.05。

（3）输出变量的含义

- h——测试的结果，其长度等于测试的次数：
  - h = 1 表示拒绝残差没有异方差的原假设。
  - h = 0 表示不拒绝残差没有异方差性零假设。即说明残差没有异方差性。
- pValue——检验统计量的概率 p 值。
- stat——检验统计量。
- cValue——检验统计量的关键值。

9.5.2 建立 ARCH 及 GARCH 模型

在 MATLAB 中建立 ARCH 及 GARCH 模型均使用函数 garch，该函数的使用方法如下。

Mdl = garch

Mdl = garch(P,Q)

Mdl = garch(Name,Value)

（1）输入变量的含义

- P——包含在 GARCH 多项式中的过去连续条件方差项的数量，指定为非负整

数。也就是说，P 是 GARCH 多项式的次数，只能使用 garch（P，Q）简写语法来指定 P，不能将 P 与 Name，Value 对参数一起指定。

如果 P> 0，那么必须将 Q 指定为正整数。

- Q——ARCH 多项式的阶数。

（2）可选择输入变量的含义

- Constant——条件方差模型的常数值。

  默认值：NaN。

- GARCH——对应于构成 GARCH 多项式的过去条件方差项的系数，指定为由‘GARCH’和非负标量的单元向量组成的逗号分隔对。如果指定 GARCHLag，则 GARCH 是与 GARCHLags 中的滞后相关的系数的等长单元向量。 否则，GARCH 是对应于滞后 1，2，…，*P* 的系数的 *P* 元素单元向量。默认情况下，GARCH 是长度为 *P*（GARCH 多项式的次数）或 numel（GARCHLags）的 NaN 的单元向量。

  例如，‘GARCH’,{0.1 0 0 0.02}。

- ARCH——对应于构成 ARCH 多项式的阶数，指定为由‘ARCH’和非负标量的单元矢量组成的逗号分隔对。如果指定 ARCHLags，则 ARCH 是与 ARCHLags 中的滞后相关的系数的等长单元向量。否则，ARCH 是对应于滞后 1，2，…，*Q* 的系数的 *Q* 元素单元矢量。默认情况下，ARCH 是长度为 *Q*（ARCH 多项式的次数）或 numel（ARCHLags）的 NaN 的单元向量。

- offset——常数项数值。

- GARCHLags——GARCH 多项式系数的滞后阶数，默认值为 1 到 *P*。

- ARCHLags——ARCH 多项式系数的滞后阶数，默认值为 1 到 *Q*。

- Distribution——模型的条件概率分布：

- Gaussian——struct(‘Name’,‘Gaussian’)。
- T——学生 t 分布

  struct(‘Name’,‘t’,‘DoF’,DoF)

  DoF 为自由度，默认为 NaN。

- Description——用已知模型描述模型内容。

（3）输出变量的含义

- Mdl——构建的条件异方差模型。

9.6 模型的应用实例

可以用 GARCH 模型对波动率进行建模，从而对未来走势进行分析。在本章中我们选用沪深 300 指数的数据进行建模及预测，建模利用如下脚本实现。

（1）读取数据

```
clc, clear all, close all
[Index,date]=xlsread('Index300','Sheet1','A2:B406');
date1 =693960+Index(:,1);
N = length(Index);
a = zeros(N,1);
for i = 1:N
   a(i) = datenum(cell2mat(date(i)));
end
a = datestr(a);
a = datevec(a);
Dates = datetime(a);
Returns = tick2ret(Index); %指数转收益率
```

（2）数据可视化

```
figure(1)
subplot(2,1,1)
plot(Dates,Index);
```

```
title('沪深 300 指数')
ylabel('指数')

subplot(2,1,2)
plot(Dates(1:end-1),Returns);
title('沪深 300 指数收益率')
ylabel('收益率')
```

如图 9-2 所示，是沪深 300 指数及收益率图像，指数有向上趋势性。

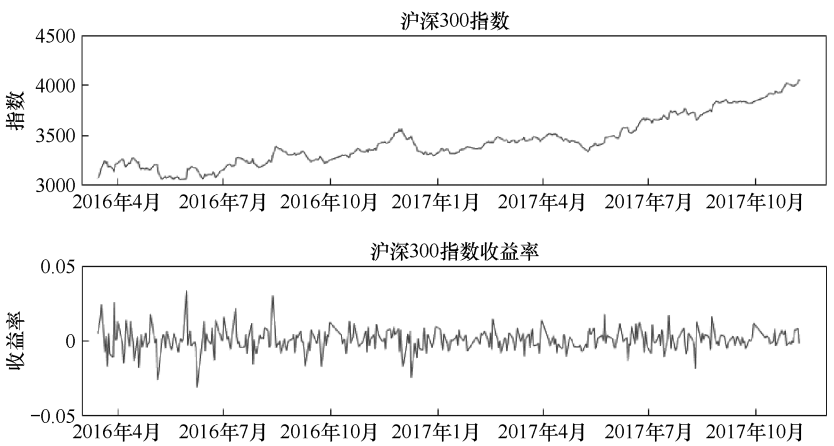


图 9-2　沪深 300 指数及收益率

（3）平稳性检验

```
disp('使用 PP 检验，如果不能拒绝原假设，则说明指数序列存在单位根')
[hp,hpValue,stat,cValue,reg] = pptest(Index,'model','TS')
if hp == 0 %存在单位根则做一阶差分
    diffIndex = diff(Index);
    [hp,hpValue,stat,cValue,reg] = pptest(diffIndex,'model','TS')
end
      [hp,hpValue,stat,cValue,reg] = adftest(diffIndex,'model','TS')
```

建立模型要先检验数据的平稳性，因此我们使用 pp 检验，数据结果如下：

```
hp =

  logical
```

```
     0


hpValue =

    0.5331
```

可以看出对训练集进行 pp 检验 hp=0，说明数据不平稳。因此，对数据进行一阶差分处理，再对差分数据进行 pp 检验：

```
hp =

  logical

   1


hpValue =

   1.0000e-03
```

对差分数据进行检验，hp=1，说明差分后的数据平稳，我们同时也可以使用 adf 检验：

```
hp =

  logical

   1


hpValue =

   1.0000e-03
```

同样可以得到差分数据平稳的结论。

（4）自相关及偏自相关函数图

```
%原指数图片
figure(2)
```

```
subplot(2,1,1)
autocorr(Index);
title('指数的自相关图像')
subplot(2,1,2)
parcorr(Index);
title('指数的偏自相关图像')
%差分后指数图片
figure(3)
subplot(2,1,1)
autocorr(diffIndex);
title('指数一阶差分后的自相关图像')
subplot(2,1,2)
parcorr(diffIndex);
title('指数一阶差分后的偏自相关图像')
```

可以通过图 9-3 看出该数据适用于 ARIMA 模型，绘制一阶差分后的自相关及偏自相关函数图如图 9-4 所示。

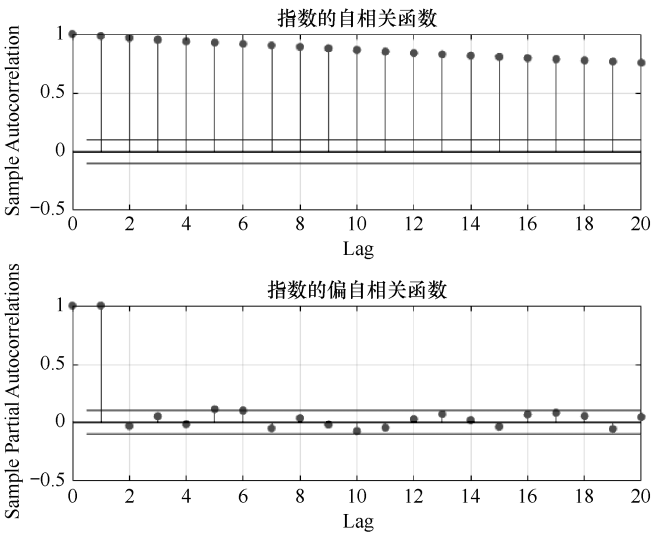


图 9-3　指数自相关及偏自相关函数图

（5）AIC 准则定阶

```
maxLags = 4;
AICSet = zeros(maxLags, maxLags);
```

```
parfor i = 1:maxLags
for j = 1:maxLags
       mdl = arima('ARLags',i,'MALags',j);
%       mdl = arima(i,1,j);
       [EstMdl, EstParamCov, logL, info] = estimate(mdl, diffIndex,
       'display', 'off');
       AICSet(i, j) = aicbic(logL, length(info.X));

end
end

% 画热度图来表示 AIC 数值的分布
figure(4)
heatmap(AICSet/1000);
xlabel('MA Lags')
ylabel('AR Lags')
title('AIC 准则图')

[OptimalARLags, OptimalMALags] = find(AICSet==min(min(AICSet)));
title(['Optimal AR and MA Lags are (' num2str(OptimalARLags) ','
num2str(OptimalMALags) ')'])
```

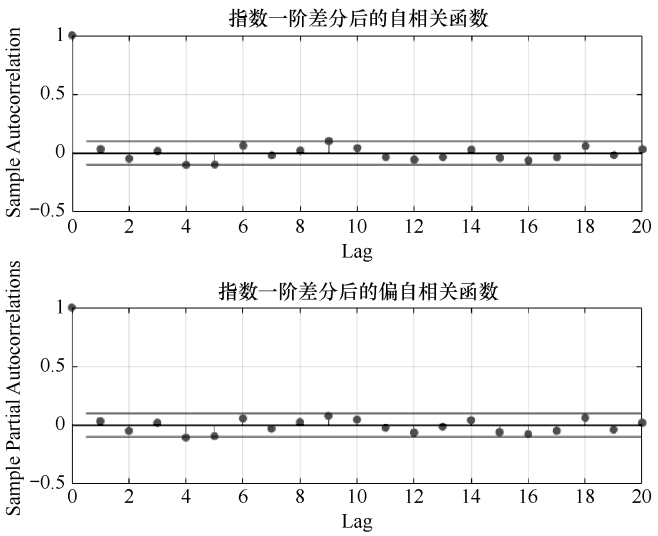


图 9-4　指数一阶差分后自相关及偏自相关函数图

为使建模更加严谨，我们使用 AIC 准则进行模型的定阶，绘制 AIC 准则的图像如图 9-5 所示，说明最适合的模型应该为 ARIMA（4,1,4）。但是对该模型进行检验，没有异方差性。为了后续的 GARCH 建模，我们建立 ARMA（4,4）模型。

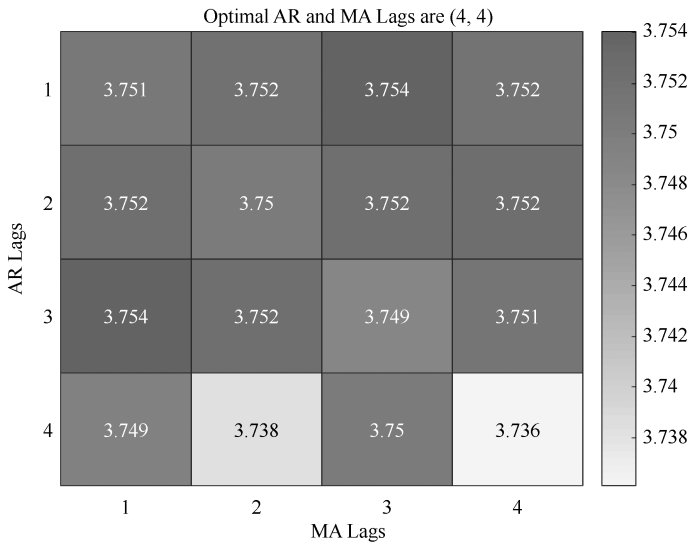


图 9-5　AIC 准则定阶分布图

（6）建立模型

```
mdl = arima('ARLags',OptimalARLags,'MALags',OptimalMALags);
fit = estimate(mdl, Index);
[res, ~, ~] = infer(fit, Index);

disp('拒绝原假设证明数据存在 ARCH 影响')
[h, pVal] = archtest(res)
```

建立 ARIMA（4,0,4）模型结果如下：

```
ARIMA(4,0,4) Model:
   --------------------
   Conditional Probability Distribution: Gaussian

                             Standard          t
    Parameter       Value        Error       Statistic
   -----------   -----------   ------------   -----------
```

```
    Constant       8.88457      32.7074      0.271638
      AR{4}             1   0.00971282       102.957
      MA{4}     -0.147354    0.0529198      -2.78448
    Variance       2354.09      144.998       16.2353
```

随后用 archtest 检验模型的异方差性，拒绝原假设证明数据存在 ARCH 影响，结果如下：

```
h =
  logical

   1

pVal =
    0
```

由于 h=1，我们将尝试建立组合的条件均值和条件方差模型。

（7）用 AIC 准则定阶 GARCH 模型

与 ARMA 参数估计过程一样，我们将对 ARCH 滞后和 GARCH 滞后计算其 AIC 准则值。

```
maxLags = 3;
AICSet = nan(maxLags, maxLags);

for p = 1:maxLags
for q = 1:maxLags
      mdl = arima('ARLags', OptimalARLags, 'MALags', OptimalMA-
      Lags, ...
'Variance', garch(p,q));
      [~,~,logL,info] = estimate(mdl, Index, 'display', 'off');
      AICSet(p,q) = aicbic(logL, length(info.X));
end
end


figure;
heatmap(AICSet/1000);
xlabel('ARCH Lags')
```

```
ylabel('GARCH Lags')
title('Akaike information criteria')

[garchLags, archLags] = find(AICSet==min(min(AICSet)));
title(['Optimal GARCH and ARCH Lags are (' num2str(garchLags) ','
num2str(archLags) ')'])
```

如图 9-6 所示，可以看到 AIC 准则值最小点为（3，2），根据 AIC 准则值建立模型。

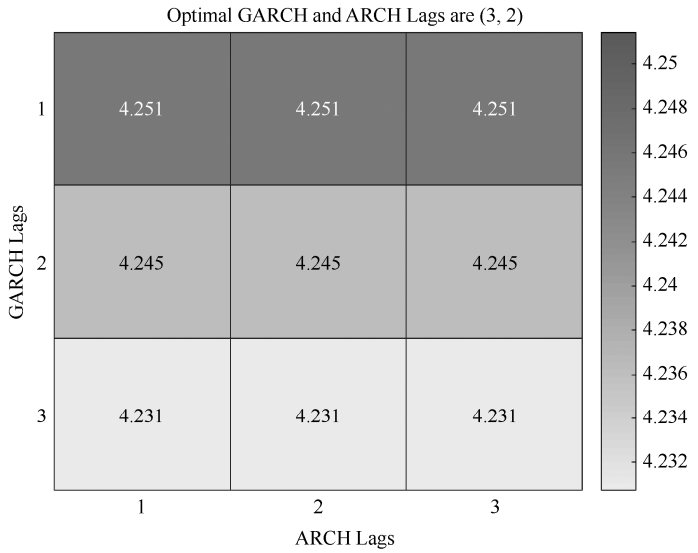


图 9-6　GARCH 模型 AIC 准则定阶分布图

（8）建立模型

```
mdl = arima('ARLags', OptimalARLags, 'MALags', OptimalMALags, ...
'Variance', garch(garchLags, archLags));
fit = estimate(mdl, Index);
```

则建立模型结果如下：

```
ARIMA(4,0,4) Model:
   --------------------
   Conditional Probability Distribution: Gaussian

                             Standard          t
    Parameter       Value        Error       Statistic
```

```
-----------   -----------   ------------   -----------
 Constant        55.786       16.249        3.43319
   AR{4}        0.98753    0.00488195        202.282
   MA{4}      -0.220016     0.0460198       -4.78091


GARCH(3,1) Conditional Variance Model:
---------------------------------------
Conditional Probability Distribution: Gaussian

                          Standard          t
 Parameter      Value         Error      Statistic
-----------   -----------   ------------   -----------
 Constant       157.895       112.29        1.40613
 GARCH{2}     0.0562699     0.102872       0.546991
 GARCH{3}      0.328563     0.121908        2.69518
  ARCH{1}      0.609481     0.124627        4.89046
```

9.7 小结

使用 ARIMA 模型拟合非平稳序列时，残差序列要求为零均值白噪声序列。如果残差序列存在异方差性，就需要建立异方差的模型。本章我们先给出了时间序列异方差性的定义及直观判断，残差序列方差齐性时，它应该在零值附近随机波动，不带任何趋势，否则就显示出异方差的性质。在 9.2 节中我们给出了异方差的检验方法，主要有 Portmantea Q 检验和 LM 检验，在 MATALB 中，使用函数 archtest 可以方便地检验异方差性。在 9.3 节中我们介绍了自回归条件异方差模型——ARCH 模型，主要是通过构造残差平方序列的自回归模型来拟合异方差函数。在 9.4 节中我们介绍了广义自回归条件异方差模型——GARCH 模型，在实践中，有些残差序列的异方差函数是具有长期自相关性的，这时如果使用 ARCH 模型拟合异方差函数，将会产生很高的移动平均阶数，因此在 GARCH 模型中加入了以前面信息为基础的预测方差项。这两个模型在 MATLAB 中用 garch 函数实现。我们在 9.5 节中给出了函数的具体使用方法，并且在 9.6 节中用沪深 300 指数的数据进行建模及预测。

参考文献

[1] 黄红梅. 应用时间序列分析. 北京：清华大学出版社，2016.

[2] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[3] Ruey S. Tsay, 王远林，王辉，等. 金融时间序列分析（第三版）. 北京：人民邮电出版社，2012.

[4] Jonathan D.Cryer, Kung-Sik Chan. Time Series Analysis With Applications in R (Second Edition): New York, Springer, 2008.

[5] Lamoureux C G , Lastrapes W D(1990). Persistence in Variance, Structural Change, and the GARCH Model[J]. Journal of Business & Economic Statistics, 8(2):225-234.

[6] 惠晓峰，柳鸿生，胡伟(2003)等. 基于时间序列 GARCH 模型的人民币汇率预测[J]. 金融研究, (5):99-105.

[7] Bollerslev T(1987). A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return[J]. Review of Economics & Statistics, 69(3):542-547.

10 多元时间序列分析

多元时间序列分析（Multivariate Time Series Analysis）是指对多变量时间序列的研究。实际中，许多问题不仅是观察单个过程 X_t，而是同时观察多个过程 X_{1t}，X_{2t},$\cdots$ X_{rt}，或者说 X_t 为向量时，需要分析多变量时间序列 $\boldsymbol{X}_t=(X_{1t},X_{2t},\cdots X_{rt})^{\mathrm{T}}$。例如，在工程上要研究电流与电压同时随时间变化的情况；在化学变化中要分析压力、温度和体积的变化关系；在气象预报分析时需要同时考虑该地区的雨量、气温和气压等记录资料。不仅要把他们的各分量看作单变量过程来研究，而且要研究各分量之间的关系及变化规律，从而对时间序列做出预报和控制，这就是多元时间序列分析，其最典型的模型就是多元 ARMA 模型。

对多元时间序列的分析很早就开始了。1976 年，Cox 和 Jenkins 在 Times Series Analysis: Forecasting and Control 一书中就采用过将天然气的输入速率作为输入变量，研究 CO_2 的输出浓度，由此将时间序列的分析领域由一元拓展到了多元的场合。

但是从技术上来讲，当时要求输入序列和研究序列都是平稳的。显然相应序列和输入序列平稳的要求是非常苛刻的，这严重限制了多元时间序列分析的运用和发展。直到 1987 年 Eagle 和 Granger 提出了协整（Cointegration）的概念。

在协整理论下，并不要求响应序列和输入序列自身平稳，只要求它们的回归残差序列平稳。残差序列平稳比响应序列与输入序列均平稳容易实现多了，这个概念的提出极大地促进了多元时间序列分析的发展，它实际上是将多元回归分析和时间序列分析有机地结合在了一起，有效地提高了预测的精度。

本章将介绍多元时间序列建模的原理和方法。

10.1 平稳多元序列建模

1987 年，Cox 和 Jenkins 采用带输入变量的 ARIMA 模型，为平稳多元序列建模。

10.1.1 平稳多元序列建模的定义

该模型的构造思想是：假设响应序列$\{Y_t\}$和输入变量序列（即自变量序列）$\{X_{1t}\}$，$\{X_{2t}\},\cdots,\{X_{kt}\}$均平稳，首先构建响应序列和输入变量序列的回归模型：

$$Y_t=\mu+\sum_{i=1}^{k}\frac{\alpha_i(L)}{\beta_i(L)}L^{l_i}X_{it}+\varepsilon_t$$

式中，$\alpha_i(L)$为第i个输入变量的自回归系数多项式；$\beta_i(L)$为第i个输入变量的移动平均系数多项式；l_i为第i个输入变量的延迟阶数；$\{\varepsilon_t\}$为回归残差序列。

因为$\{Y_t\}$和$\{X_{1t}\},\{X_{2t}\},\cdots,\{X_{kt}\}$均平稳，平稳序列的线性组合仍然是平稳的，所以残差序列$\{\varepsilon_t\}$为平稳序列：

$$\varepsilon_t=Y_t-\left(\mu+\sum_{i=1}^{k}\frac{\alpha_i(L)}{\beta_i(L)}L^{l_i}X_{it}\right)$$

使用 ARMA 模型继续提取残差序列$\{\varepsilon_t\}$中的相关信息。最终得到的模型为：

$$\begin{cases}Y_t=\mu+\sum_{i=1}^{k}\frac{\alpha_i(L)}{\beta_i(L)}L^{l_i}X_{it}+\varepsilon_t\\ \varepsilon_t=\frac{\alpha(L)}{\beta(L)}a_t\end{cases}$$

模型被称为动态回归模型，简记为 ARIMAX。式中，$\alpha(L)$为残差序列自回归系数多项式；$\beta(L)$为残差序列移动平均系数多项式；a_t为零均值白噪声序列。

10.1.2 虚假回归

在 ARIMAX 模型中，如果平稳性条件不满足，就容易产生虚假回归的问题。我们考虑一元线性回归模型$Y_t=\beta_0+\beta_1X_t+\varepsilon_t$，为了检验模型的显著性，就需要对模拟模型进行检验：

$$\mathrm{H}_0:\beta_1=0\leftrightarrow\mathrm{H}_1:\beta_1\neq0$$

假定响应序列和输入变量序列相互独立，理论上，检验结果应该接受原假设H_0，如检验结果为拒绝H_0，那么我们就接受了一个本不应该成立的回归模型，从而犯了

第一类错误（纳伪错误）。由于样本的随机性，故纳伪错误会一直存在，我们可采用显著性水平 α 控制纳伪错误的概率 $P\left(\mathrm{H}_1\middle|\mathrm{H}_0\right)=\alpha$ 。

通常情况下，我们采用 t 统计量进行参数显著性检验：$t=\dfrac{\beta_1}{\alpha_\beta}$ 。当 $\{Y_t\}$ 和 $\{X_t\}$ 都平稳时，该统计量服从自由度为样本容量的 t 分布，从而有：

$$P\left(|t|\leqslant t_{\alpha/2}(n)\text{平稳序列}\right)\leqslant\alpha$$

即当 $|t|\leqslant t_{\alpha/2}(n)$ 时，纳伪错误精确控制在 α 内。

当 $\{Y_t\}$ 和 $\{X_t\}$ 不平稳时，随机模拟的结果显示，检验统计量 $t=\dfrac{\beta_1}{\alpha_\beta}$ 不服从 t 分布，

这时 t 统计量的样本方差远远大于 t 分布的方差，如果仍采用 t 分布的临界值进行检验，拒绝原假设的概率就会大大增加，这样导致我们无法控制纳伪错误，容易接受本不该成立的回归模型，这种现象称为虚假回归。

10.2 协整

Eagle（恩格尔）与 Granger（格兰杰）1978 年首次提出协整的概念，并将经济变量之间存在的长期稳定关系称为“协整关系”。协整理论分析时间序列的非平稳经济变量间蕴含的长期稳定关系，从而为协整变量之间建立误差修正模型奠定了理论基础。

协整概念及其方法的提出对于用非平稳变量建立经济计量模型非常重要。当且仅当若干个非平稳变量具有协整关系时，由这些变量建立的回归模型才有意义，所以协整性检验也是区别真实回归和虚假回归的有效方法。

10.2.1 单整的概念

假如原序列一阶差分后平稳，说明序列存在一个单位根，这时称序列为一阶单整序列，简记为 $I(0)$ 。

假如原序列至少需要进行 d 阶差分才能实现平稳，说明原序列存在 d 个单位根，这时称原序列为 d 阶单整序列简记为 $I(d)$

若 $X_t : I(0)$，对任意非零实数 a,b，有 $\mathrm{a}+\mathrm{b}X_t : I(0)$；

若 $X_t : I(\mathrm{d})$，对任意非零实数 a,b，有 $\mathrm{a}+\mathrm{b}X_t : I(\mathrm{d})$；

若 $X_t : I(0)$，$Y_t : I(0)$ 对任意非零实数 a,b，有 $Z_t = \mathrm{a}X_t + \mathrm{b}Y_t : I(0)$；

若 $X_t : I(\mathrm{d})$，$Y_t : I(\mathrm{c})$ 对任意非零实数 a,b，有 $Z_t = \mathrm{a}X_t + \mathrm{b}Y_t : I(k)\ k \leqslant \max[\mathrm{d},\mathrm{c}]$

10.2.2 协整的概念

在现实生活中我们会发现，有些序列自身的变化虽然是非平稳的，但是序列与序列之间却具有非常密切的长期均衡关系。

比如说家庭人均纯收入和人均生活消费支出，单整分析显示两个序列是非平稳的，但是将这两个序列联合起来考虑，通过观察它们的时序图，我们却发现他们之间具有非常稳定的线性相关关系。当收入增多时，生活消费支出也增多，它们的变化速度几乎一致。这种稳定的同变关系让我们怀疑它们之间的具有一种内在的平稳机制，导致了它们自身的变化虽然是不平稳的，但是彼此之间却具有长期均衡关系。如图 10-1 所示是城镇居民人均可支配收入及支出，单个看每个数据都是非平稳的，但是两者间存在非常稳定的关系。

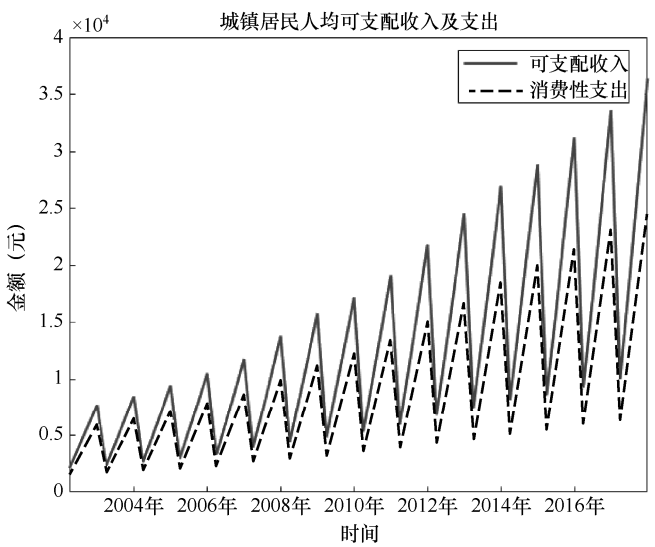


图 10-1　城镇居民人均可支配收入及支出

假定自变量序列为$\{X_1\},\{X_2\},\cdots,\{X_k\}$，响应变量序列为$\{Y_t\}$，构造回归模型：

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{it} + \varepsilon_t$$

假定回归残差序列$\{\varepsilon_t\}$平稳，我们称响应序列$\{Y_t\}$与子变量序列$\{X_1\},\{X_2\},\cdots,\{X_k\}$之间具有协整关系。

协整概念的提出有着非常重要的意义，因为我们之前一直不敢大胆地对非平稳序列构建动态回归模型，是担心非平稳序列容易产生虚假回归的问题。而虚假回归之所以会产生是因为残差序列不平稳。如果非平稳序列之间具有协整关系，那就说明残差序列平稳，那就不会产生虚假回归问题了。

这说明要将输入变量引入响应序列建模，不一定要所有的序列都平稳，只需要它们之间具有协整关系。这个限制条件显然比 Cox 和 Jenkins 要求所有序列都平稳的限制条件要宽松很多，这极大地拓宽了动态回归模型的适用范围。

（1）协整只涉及非平稳变量的线性组合。从理论上而言，在一组非平稳变量中，极有可能存在着非线性的长期均衡关系。

（2）协整只涉及阶数相同的单整变量。如果变量的单整阶数不同，则按照通常的学术意义，可以认为它们不存在协整关系。

（3）大多数协整的相关研究集中在每个变量只有一个单位根的情况，而极少数的经济变量是单整阶数大于 1 的变量。

10.2.3 协整检验

多元非平稳序列之间能否建立动态回归模型，关键在于它们之间是否具有协整关系。所以要对多元非平稳序列建模必须得先进行协整检验，也称为 Eagle-Granger 检验，简称为 EG 检验。

（1）假设条件

由于自然界中绝大多数序列之间不具有协整关系，所以 EG 检验的假设条件如下确定。

H_0：多元非平稳序列之间不存在协整关系。

H_1：多元非平稳序列之间存在协整关系。

由于协整关系主要是通过考察回归残差的平稳性确定，所以上述假设条件等价于：

H_0：回归残差序列$\{\varepsilon_t\}$非平稳。

H_1：回归残差序列$\{\varepsilon_t\}$平稳。

（2）检验步骤

EG 检验也称为 EG 两步法，它遵循如下两个步骤进行。

步骤一：建立响应序列与输入序列之间的回归模型：

$$Y_t = \hat{\beta}_0 + \hat{\beta}_1 X_{1t} + \cdots + \hat{\beta}_k X_{kt} + \varepsilon_t$$

式中，$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$是最小二乘估计值。

步骤二：对回归残差序列$\{\varepsilon_t\}$进行平稳性检验。

我们主要是采用单位根检验的方法来考察回归残差序列的平稳性，所以，假设条件等价于：

$$H_0 : \varepsilon_t : I(k), k \geqslant 1 \leftrightarrow H_1 : \varepsilon_t : I(0)$$

10.3 模型的 MATLAB 方法

在 MATLAB 中，EG 检验通过函数 egcitest 实现，具体使用方法如下：

[h,pValue,stat,cValue,reg1,reg2] = egcitest(Y)

[h,pValue,stat,cValue,reg1,reg2] = egcitest(Y,Name,Value)

（1）输入参数含义

- Y——时间序列数据。行数为观测样本数，列数为特征数，特征数不大于 12。

（2）可选择的输入参数含义

- creg——字符向量，表示协整回归的形式， $y_1 = Y(:,1), y_2 = Y(:,2:end)$, 并且确定自变量选项，默认为‘c’：

  $y_1 = X\,\mathrm{a} + y_2\,\mathrm{b} + \varepsilon$

  - 'nc'表示 X 中没有固定项或者趋势项。
  - 'c'表示 X 中有常数项，但是没趋势项。
  - 'ct'表示 X 中有常数项及线性趋势项。
  - 'ct'表示 X 中有常数项、线性项及二次趋势项。
- cvec——协整回归中系数向量[a;b],a 的长度为 0，1，2 或者 3，取决于 creg 的输入，b 的长度为特征数−1，并且单位化。
- rreg——残量回归的形式，默认为'ADF'。
  - 'ADF'——增广的 DF 检验。
  - 'PP'——PP 检验。
- lags——残量回归中的阶数，默认为 0。
- test——残量回归的检验统计量，默认为 t1。
  - 't1'——t 检验统计量。
  - 't2'——z 检验统计量。
- alpha——显著性水平，默认为 0.05。

（3）输出变量的含义

- h——测试的布尔决策向量，长度等于测试的次数。值等于 1 则表示拒绝原假设，数据有协整关系。值等于 0 则表示不能拒绝原假设，数据没有协整关系。
- pValue——检验统计量的 p 值。
- stat——检验统计量，长度等于检验的次数。
- cValue——检验的关键值向量，长度等于检验的数量。值取左尾概率。

- reg1——协整回归的回归统计量结果。
- reg2——残量回归的回归统计结果，有如下内容：
  - num——时间序列长度。
  - size——有效的样本量，根据滞后数调整。
  - names——回归系数名。
  - coeff——估计的系数值。
  - se——估计系数的标准差。
  - cov——估计系数的方差矩阵。
  - tStats——t 统计量的系数及概率 p 值。
  - FStat——F 统计量及 p 值。
  - yMu——输入序列调整滞后数的均值。
  - ySigma——输入序列调整滞后数的标准差。
  - yHat——输入序列调整滞后数的适应值。
  - res——回归残差项。
  - DWStat——Durbin-Watson 统计量。
  - SSR——回归平方和。
  - SSE——误差平方和。
  - SST——总平方和。
  - MSE——均方误差。
  - RMSE——回归的标准误差。
  - RSq——R^2 统计量。
  - aRSq——调整后的 R^2 统计量。

- LL——高斯似然数。
- AIC——AIC 值。
- BIC——BIC 值。
- HQC——Hannan-Quinn 信息准则。

10.4 模型的应用实例

我们使用 2002—2017 年城镇居民可支配收入及支出来建立模型。数据如图 10-1 所示，具体的实现脚本如下。

（1）读取数据

```
clc, clear all, close all
[Y,date] = xlsread('mul_data','Sheet1','A2:C65');
date1 = 693960+Y(:,1);
N = length(Y);
a = zeros(N,1);
for i = 1:N
   a(i) = datenum(cell2mat(date(i)));
end
a = datestr(a);
a = datevec(a);
Dates = datetime(a);
```

（2）数据可视化

```
figure(1)
plot(Dates,Y(:,1),'LineWidth',1.5);
hold on
plot(Dates,Y(:,2),'r','LineWidth',1.5);
title('城镇居民人均可支配收入及支出')
ylabel('金额(元）')
xlabel('时间')
legend('可支配收入','消费性支出')
```

如图 10-1 所示，很明显单个的数据都是不平稳的，但是我们对其做协整检验。

（3）协整检验

```
[h,pValue,stat,cValue,reg] = egcitest(Y);
a = reg.coeff(1);
b = reg.coeff(2);
figure(2)
plot(Dates,Y*[1;-b]-a,'LineWidth',1.5)
title('城镇居民人均可支配收入及支出协整关系图')
```

协整检验的结果如下：

```
h =

  logical

   1

pValue =

   1.0000e-03
```

检验结果 h=1，则说明这两个数据之间存在协整关系，画协整关系 $y_1 - X\,\mathrm{a} - y_2\mathrm{b}$ 的图如图 10-2 所示。

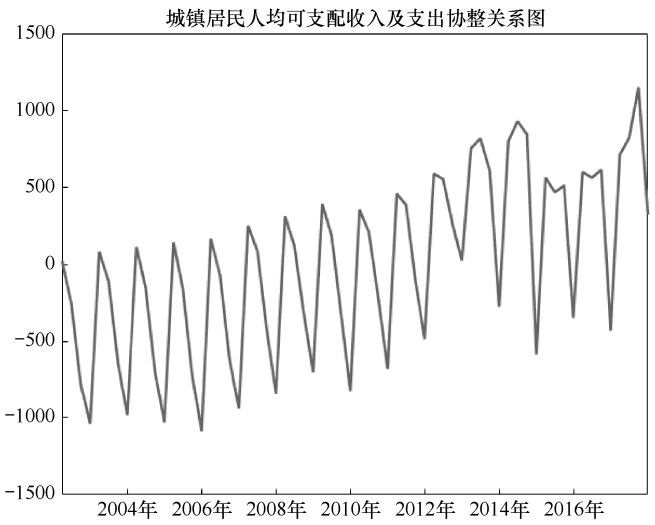


图 10-2　城镇居民人均可支配收入及支出协整关系图

（4）残差序列 pp 检验

```
coin_res = Y*[1;-b]-a;
disp('使用 PP 检验，如果不能拒绝原假设，则说明指数序列存在单位根')
[hp,hpValue,stat,cValue,reg] = pptest(coin_res,'model','TS')
```

我们对残差数据进行 pp 检验，可得：

```
hp =

  logical

   1



hpValue =

   1.0000e-03
```

由此可知残差序列是平稳的。

（5）AIC 定阶

```
maxLags = 4;
AICSet = zeros(maxLags, maxLags);

for i = 1:maxLags
for j = 1:maxLags
       mdl = arima('ARLags',[1:i],'MALags',[1:j]);
       [EstMdl, EstParamCov, logL, info] = estimate(mdl, coin_res,
       'display', 'off');
       AICSet(i, j) = aicbic(logL, length(info.X));

end
end

% 画热度图来表示 AIC 数值的分布
figure(3)
```

```
heatmap(AICSet/1000);
xlabel('MA Lags')
ylabel('AR Lags')
title('AIC 准则图')

[OptimalARLags, OptimalMALags] = find(AICSet==min(min(AICSet)));
```

对其进行 AIC 准则定阶可知最优阶数为（4，2），如图 10-3 所示。

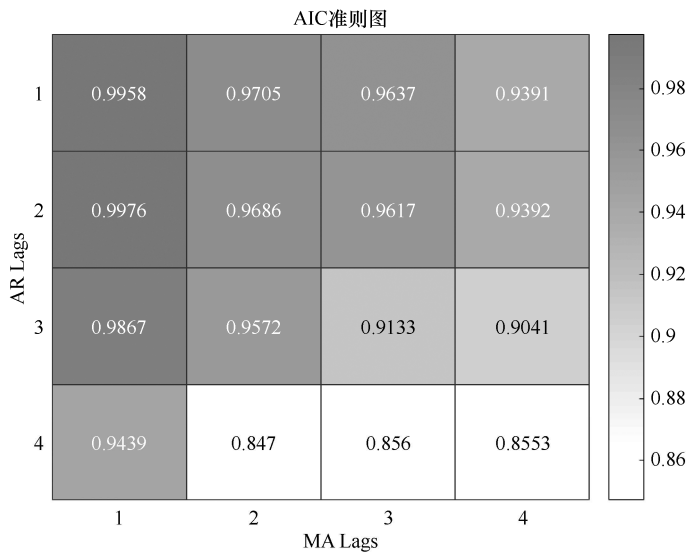


图 10-3　AIC 准则定阶分布图

（6）建立模型

```
mdl = arima( OptimalARLags,0, OptimalMALags);
fit = estimate(mdl, coin_res);
```

对残差数据建立 ARMA（4,2）模型可得：

```
 ARIMA(4,0,2) Model:
    --------------------
    Conditional Probability Distribution: Gaussian

                               Standard          t
     Parameter       Value         Error       Statistic
```

```
-----------   -----------   ------------   -----------
 Constant       44.7683       54.2852       0.824688
   AR{1}      0.0244514     0.0539516       0.453211
   AR{2}      0.0226033     0.0552576       0.409054
   AR{3}     -0.0197701     0.0405411      -0.487655
   AR{4}       0.972715     0.0536397        18.1342
   MA{1}       0.526393      0.121206        4.34298
   MA{2}       0.224006      0.121025         1.8509
 Variance       25516.6       3513.27        7.26293
```

（7）模型预测

我们将数据按 70%及 30%的比例分为训练集及预测集，用样本内数据建立模型。我们取前 1 至 54 个数据为训练集，用该模型预测后 10 个残差时间序列的值，再利用两个特征的协整关系，根据消费性支出计算出居民人均可支配收入。具体过程如下：

```
trg = round(0.85 * N);
TrainingData = coin_res(1:trg); % 前 70%
TestData = coin_res(trg+1:end); % 后 30%

y2 = Y(1:trg,2);
y2_pre = Y(trg+1:end,2);
y1 = Y(1:trg,1);
[Yf,YMSE] = forecast(fit,10,'Y0',coin_res(1:trg));
y1_pre = y2_pre*b+a+Yf;

figure(4)
plot(Y(:,1),'b','LineWidth',2)
hold on
plot(trg+1:trg+10,y1_pre,'r','LineWidth',1);
title('模型预测情况')
legend('实际数据','预测值')
hold off
```

如图 10-4 所示，我们建立模型预测残差序列值，再利用协整关系建模预测未来值，可以看到红色的预测值与蓝色的实际数据间误差极小。

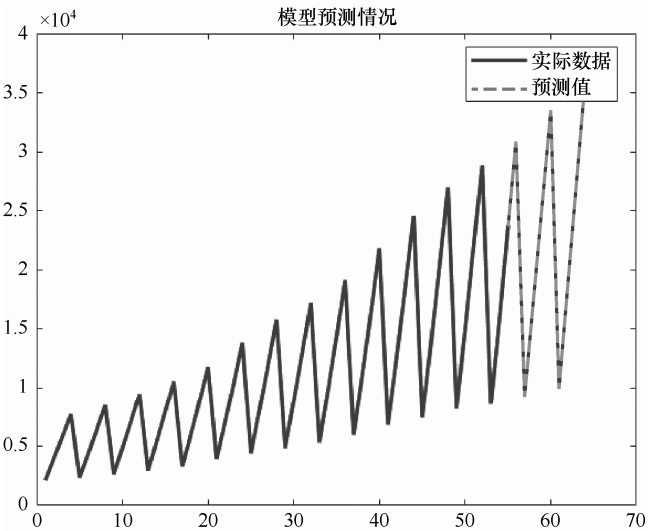


图 10-4　模型预测情况图

10.5　小结

在实际中，许多问题不仅需要观察单个过程，而需要同时观察多个过程，因此需要使用多元时间序列模型。在本章中我们首先介绍了平稳多元序列建模的方法。但是对于多元序列来说，由于存在多组数据，在现实生活中我们会发现，有些序列自身的变化虽然是非平稳的，但是序列与序列之间却具有非常密切的长期均衡关系。因此在 10.2 节中我们介绍了协整的概念，如果非平稳序列之间具有协整关系，那就说明残差序列平稳，我们可以建立时间序列模型。对于协整关系可以使用 Eagle-Granger 检验，该检验可以在 MATLAB 中使用 egcitest 函数实现。我们在 10.3 节中对该函数的使用方法进行了详细的解释。在 10.4 节中，我们使用 2002—2017 年城镇居民可支配收入及支出数据来建立模型。

参考文献

[1] Javed W, Mcdonnel B, Elmqvist N(2010). Graphical Perception of Multiple Time

Series[M]. IEEE Educational Activities Department.

[2] Widiputra H, Pears R, Kasabov N. Multiple Time-Series Prediction through Multiple Time-Series Relationships Profiling and Clustered Recurring Trends[C]// Advances in Knowledge Discovery and Data Mining -, Pacific-Asia Conference, PAKDD 2011, Shenzhen, China, May 24-27, 2011, Proceedings. DBLP, 2011:161-172.

[3] Tiao G C, Box G E P. Modeling Multiple Times Series with Applications[J]. Journal of the American Statistical Association, 1981, 76(376):802-816.

[4] 王燕. 应用时间序列分析（第三版）. 北京：中国人民大学出版社，2012.

[5] Ruey S. Tsay, 王远林，王辉，等. 金融时间序列分析（第三版）. 北京：人民邮电出版社，2012.

[6] Jonathan D.Cryer, Kung-Sik Chan. Time Series Analysis With Applications in R (Second Edition): New York, Springer, 2008.

11 航空公司乘客预测的时间序列模型

前面章节侧重介绍各个时间序列模型的特征及方法，对于一组陌生数据来说，如何来评判是否可用时间序列方法来建模？如果可以，如何一步步实现整个建模过程？本章将通过一个实例展示如何使用 MATLAB 来实现具体的时间序列分析和建模过程。

本章所研究的对象是航空公司每月的旅客数据，希望通过历史数据来预测未来的旅客数，从而为航空公司的未来发展规划提供决策支持。这个案例的特点是数据具有明显的时序特征，所以可以很容易想到用时间序列方法进行建模。在接下来的章节中，将按照时间分析的标准流程 Box-Jenkins 来一步步解决整个问题。

11.1 时序数据的分析

11.1.1 加载航空公司乘客数据

加载数据，然后绘制每月乘客数的自然对数（时间序列是 1949—1960 年每月国际航空旅客人数）。

```
clc, clear, close all
load Data_Airline
Y = log(Dataset.PSSG);
N = length(Y);
mos = get(Dataset,'ObsNames');

figure(1)
plot(Y); xlim([1,N])
set(gca,'XTick',[1:18:N])
set(gca,'XTickLabel',mos([1:18:N]))
title('航空公司乘客数的自然对数')
ylabel('乘客数(千)')
xlabel('时间(月/年)')
```

运行本节程序，得到如图 11-1 所示的趋势图。从图中可以看出数据不稳定，但有明显的线性趋势和季节周期性。

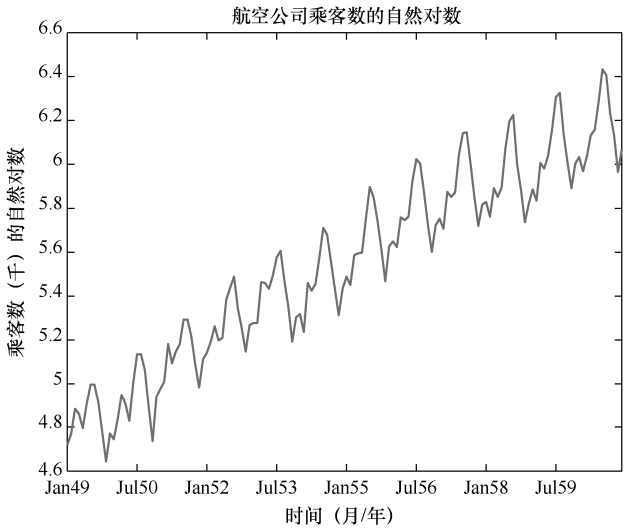


图 11-1　乘客数（自然对数）的时间序列趋势图

11.1.2　差分序列分析

首先，用 $\Delta y_t = (1-L)(1-L^{12})y_t$ 计算差分序列。

其中，y_t 为对原始数据进行对数转化后的数据。

然后，对这个差分序列进行数据可视化。

```
A1 = LagOp({1,-1},'Lags',[0,1]);
A12 = LagOp({1,-1},'Lags',[1,12]);
dY = filter(A1*A12,Y);

figure(2)
plot(dY)
title('航空公司乘客数的差分序列')
ylabel('乘客数(千)')
xlabel('差分序列编号')
```

执行该节脚本可以得到如图 11-2 所示的该序列的差分序列图，从图中可以看出，差分序列具有很好的稳定性。

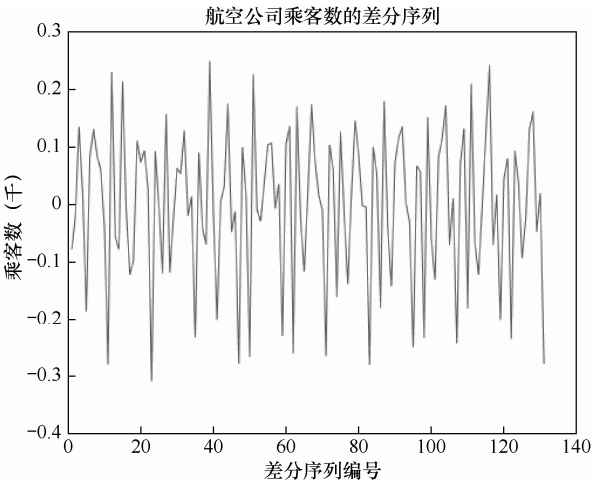


图 11-2　乘客序列差分序列图

11.1.3　自相关函数（ACF）分析

绘制差分序列的自相关函数。

```
figure(3)
autocorr(dY,50)
```

执行本节脚本可以得到如图 11-3 所示的自相关函数图。样本的自相关函数显示 12 次差分序列具有非常大的自相关性，差分次数少时同样具有潜在的自相关性。

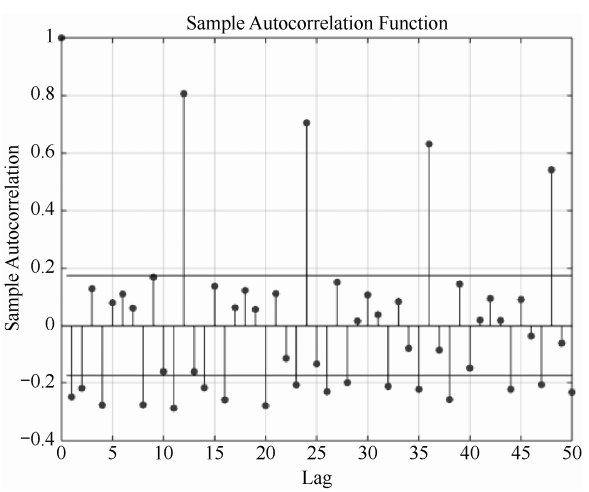


图 11-3　样本的自相关函数图

11.2 模型的估计

11.2.1 定义季节 ARIMA 模型

在对数据进行分析后， 已大致确定这个时序数据的特点。根据 Box and Jenkins 建议的季节模型：$(1-L)(1-L^{12})y_t=(1-\theta_1)(1-\theta_{12})\varepsilon_t$ 就可以先指定这个模型。

```
model = arima('Constant',0,'D',1,'Seasonality',12,...
            'MALags',1,'SMALags',12)
model =
  arima - 属性:
     Description: "ARIMA(0,1,1) Model Seasonally Integrated with
     Seasonal MA(12) (Gaussian Distribution)"
     Distribution: Name = "Gaussian"
     P: 13
     D: 1
     Q: 13
     Constant: 0
     AR: {}
     SAR: {}
     MA: {NaN} at lag [1]
     SMA: {NaN} at lag [12]
     Seasonality: 12
     Beta: [1×0]
     Variance: NaN
```

程序运行的结果显示，参数 P 等于 13，对应非季节性和季节性差异度（1 + 12）的总和；属性 Q 也等于 13，对应非季节性和季节性的 MA 多项式（1 + 12）的和，需要估计的参数值为 NaN，也就是说没有额外需要估计的参数了。

11.2.2 使用样本数据估计模型

用前面 13 个观测数据作为初样数据，余下的 131 个观测作为估计。

```
Y0 = Y(1:13);
```

```
[fit,VarCov] = estimate(model,Y(14:end),'Y0',Y0);
display(fit)
print(fit, VarCov)
    ARIMA(0,1,1) Model Seasonally Integrated with Seasonal MA(12)
(Gaussian Distribution):

      Value      StandardError    TStatistic      PValue
    _________    _____________    __________    __________

   Constant        0              0           NaN           NaN
   MA{1}     -0.37716       0.073426     -5.1366      2.7972e-07
   SMA{12}   -0.57238       0.093933     -6.0935      1.1047e-09
   Variance  0.0013887    0.00015242      9.1115      8.1249e-20
fit =
  arima - 属性:
    Description: "ARIMA(0,1,1) Model Seasonally Integrated with
                  Seasonal MA(12) (Gaussian Distribution)"
   Distribution: Name = "Gaussian"
              P: 13
              D: 1
              Q: 13
       Constant: 0
             AR: {}
            SAR: {}
             MA: {-0.377161} at lag [1]
            SMA: {-0.572379} at lag [12]
  Seasonality: 12
           Beta: [1×0]
       Variance: 0.00138874
   ARIMA(0,1,1) Model Seasonally Integrated with Seasonal MA(12):
   ----------------------------------------------------------------
   Conditional Probability Distribution: Gaussian

                              Standard        t
   Parameter      Value         Error      Statistic
   -----------   -----------   ------------   -----------
   Constant            0        Fixed         Fixed
```

```
      MA{1}      -0.377161     0.0734258       -5.13662
    SMA{12}      -0.572379     0.0939327        -6.0935
   Variance     0.00138874   0.000152417          9.1115
```

11.3 模型的测试

11.3.1 计算残差

计算模型的残差，结果如图 11-4 所示。

```
res = infer(fit,Y(14:end),'Y0',Y0);
figure(4)
plot(14:N,res)
ylabel('时序残差(千)')
xlabel('时序编号')
xlim([0,N])
title('残差')
% When you use the first 13 observations as presample data, residuals
are
% are available from time 14 onward
```

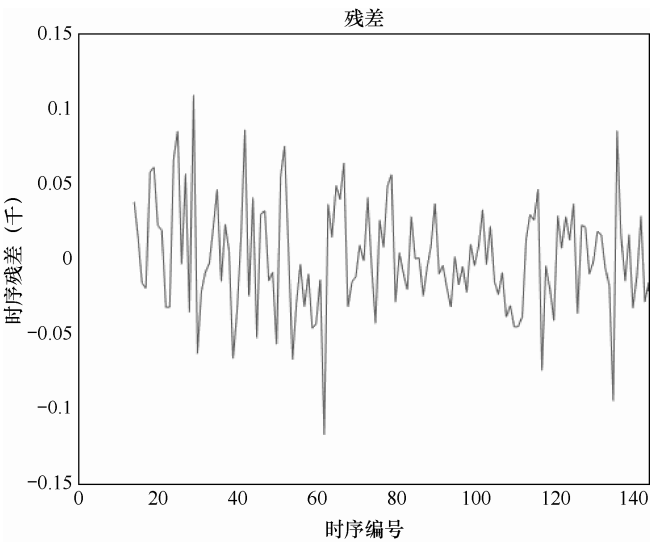


图 11-4　乘客序列的残差图

11.3.2 检查残差的分布

模型的一个假设是残差遵循高斯分布，因此可以计算模型的残差，然后检查残差的分布，如图 11-5 所示。

```
res = infer(fit,Y);
stres = res/sqrt(fit.Variance);

figure(5)
subplot(1,2,1)
qqplot(stres)

x = -4:.05:4;
[f,xi] = ksdensity(stres);
subplot(1,2,2)
plot(xi,f,'k','LineWidth',2); hold on
plot(x,normpdf(x),'r--','LineWidth',2); hold off
legend('标准化残差','标准正态')
```

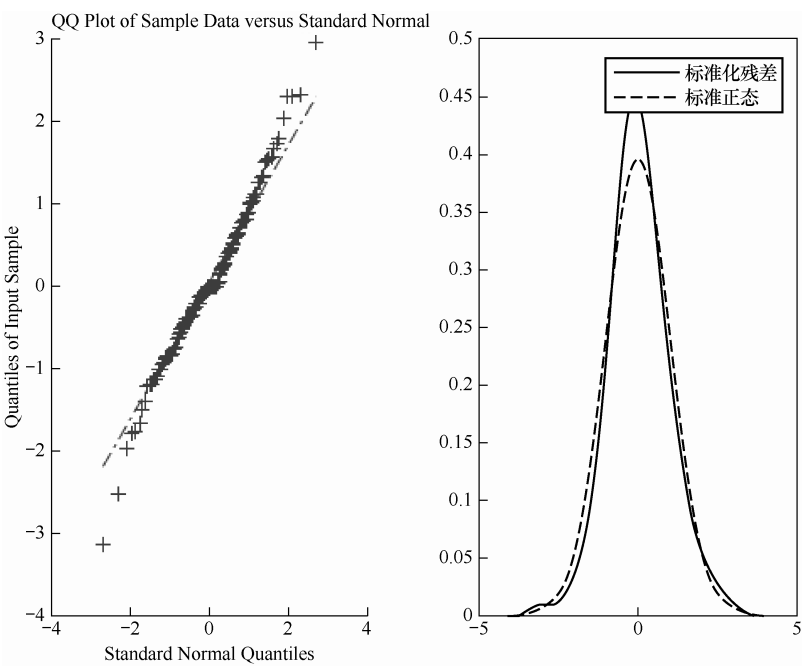


图 11-5 乘客序列的残差分布图

11.3.3 检查自相关性

确认残差是不相关的，然后查看样本自相关函数（ACF）和部分自相关函数（PACF）的标准化残差，如图 11-6 所示。

```
figure(6)
subplot(2,1,1)
autocorr(stres)
subplot(2,1,2)
parcorr(stres)
```

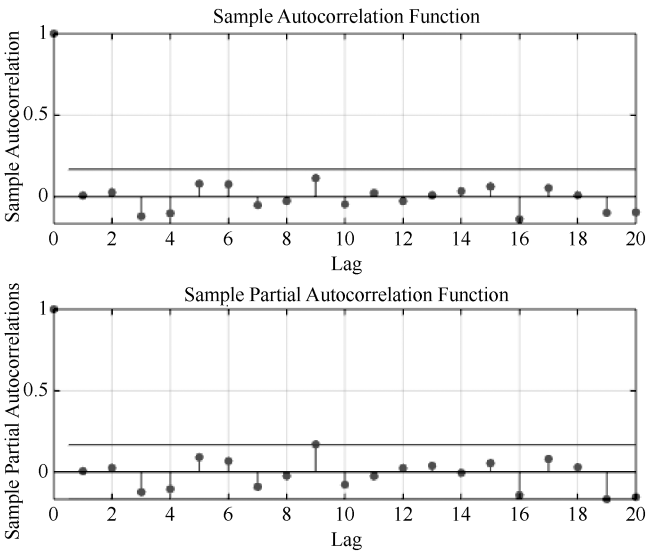


图 11-6　乘客序列的残差分布图

样本 ACF 和 PACF 图显示此序列没有显著的自相关。更正式的是，Ljung-Box Q 测试（LBQ，是对 Randomness 的检验，或者说是对时间序列是否存在滞后相关的一种统计检验）进行延迟 5、10 和 15 的测试，自由度分别为 3、8 和 13。自由度解释了两个估计的移动平均系数。LBQ 测试确认了样本 ACF 和 PACF 结果，所有的自相关系数都等于零的假设，在三个滞后的情况下都不会被拒绝（$h = 0$）。

```
[h,p] = lbqtest(stres,'lags',[5,10,15],'dof',[3,8,13])
 h =
  1×3 logical 数组
```

```
     0   0   0
p =
    0.1842    0.3835    0.7321
```

11.3.4 检查预测的表现

使用“holdout”保留样例来计算模型的预测 MSE。使用前 100 个观察值来估计模型，然后预测接下来的 44 个周期。预测值与真实数值的比较（预测误差）如图 11-7 所示。

```
Y1 = Y(1:100);
Y2 = Y(101:end);

fit1 = estimate(model,Y1);
Yf1 = forecast(fit1,44,'Y0',Y1);
PMSE = mean((Y2-Yf1).^2)

clf
plot(Y2,'r','LineWidth',2)
hold on
plot(Yf1,'k--','LineWidth',1.5)
xlim([0,44])
title('预测误差')
legend('观察','预测','Location','NorthWest')
hold off

   ARIMA(0,1,1) Model Seasonally Integrated with Seasonal MA(12)
   (Gaussian Distribution):

              Value      StandardError    TStatistic      PValue
            _________    _____________    __________    __________

   Constant            0              0         NaN             NaN
   MA{1}       -0.35674       0.089461     -3.9876      6.6739e-05
   SMA{12}     -0.63319       0.098744     -6.4124      1.4326e-10
   Variance   0.0013285    0.00015882       8.365       6.013e-17
PMSE =
   0.0069
```

图 11-7 是预测值与实际数值的比较，从图中可以看出模型的预测能力相当好，也可以对 PMSE 与 MMSE 进行比较，以获得更好的模型。

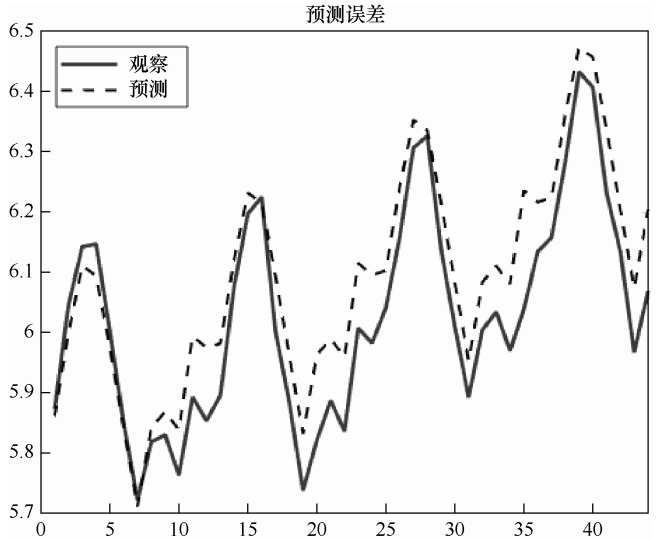


图 11-7　预测值与实际数值的比较

11.4　模型预测

11.4.1　预测航空公司乘客数

使用拟合的模型在一个 60 个月（5 年）的时间范围内生成 MMSE 预测和相应的均方误差。使用观察时间序列作为初样数据。在默认情况下，使用指定的模型来预测这些初样数据的预测值。

```
[Yf,YMSE] = forecast(fit,60,'Y0',Y);
upper = Yf + 1.96*sqrt(YMSE);
lower = Yf - 1.96*sqrt(YMSE);

figure
plot(Y,'Color',[.75,.75,.75])
hold on
h1 = plot(N+1:N+60,Yf,'r','LineWidth',2);
```

```
h2 = plot(N+1:N+60,upper,'k--','LineWidth',1.5);
plot(N+1:N+60,lower,'k--','LineWidth',1.5)
xlim([0,N+60])
title('预测和 95%的预测置信区间')
legend([h1,h2],'预测','95% 置信区间','Location','NorthWest')
hold off

% The MMSE forecast shows airline passenger counts continuing to
% grow over the forecast horizon. The confidence bounds show that a
  decline in
% passenger counts is plausible, however. Because this is a
% nonstationary process, the width of the forecast intervals grows
  over time.
```

运行本节程序，会得到如图 11-8 所示的预测置信区间，通过置信区间可以判断预测的准确程度。

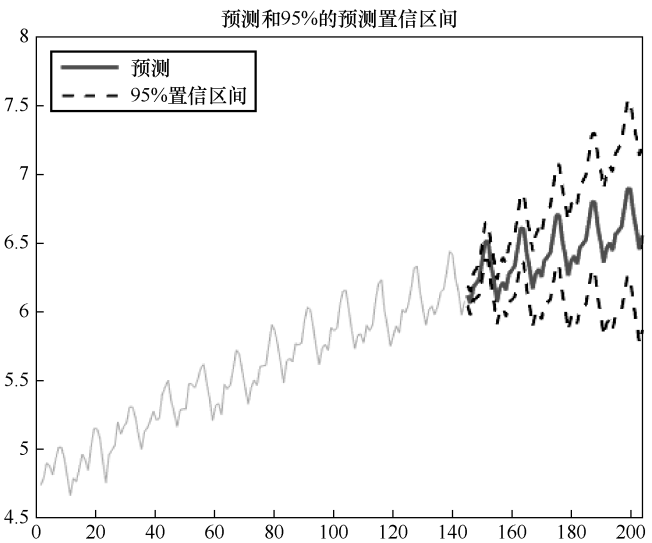


图 11-8　预测的置信区间

11.4.2　仿真航空公司乘客数

使用拟合的模型对航空公司 60 个月（5 年）的乘客数进行 25 次仿真，并绘制仿真结果。

```
rng('default')
Ysim = simulate(fit,60,'numPaths',25,'Y0',Y,'E0',res);
mn = mean(Ysim,2);

figure
plot(Y,'k')
hold on
plot(N+1:N+60,Ysim,'Color',[.85,.85,.85]);
h = plot(N+1:N+60,mn,'k--','LineWidth',2);
xlim([0,N+60])
title('模拟的航空公司乘客数量')
legend(h,'模拟均值','Location','NorthWest')
hold off

% The simulated forecasts show growth and seasonal periodicity
similar to
% the observed series.
```

运行本节程序，会得到如图 11-9 所示的蒙特卡罗仿真结果，可以看出经过多次仿真后，预测结果的总体确实不变，且分布相对集中，这说明预测的结果具有较高的可信度。

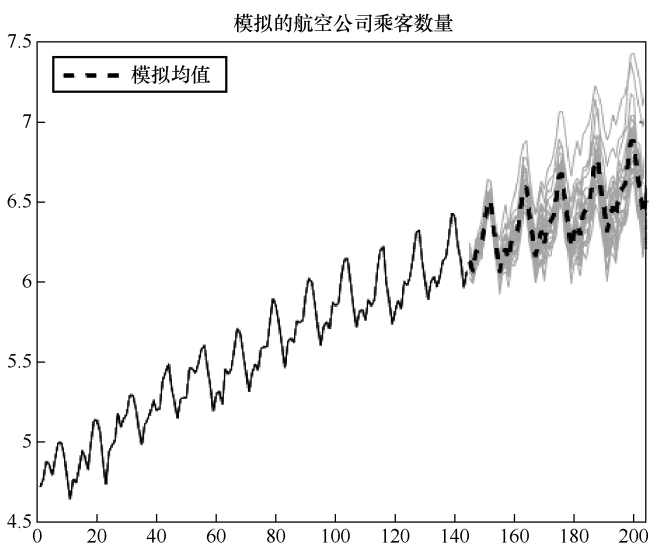


图 11-9　蒙特卡罗仿真结果分布

11.5 模型的评估

11.5.1 估计未来事件的概率

使用模拟来估计在未来 5 年的某个时间，航空公司乘客数量将达到或超过某个值的概率，使用如下脚本来计算与估计误差。

```
rng('default')
Ysim = simulate(fit,60,'numPaths',1000,'Y0',Y,'E0',res);

g7 = sum(Ysim >= 7) > 0;
phat = mean(g7)
err = sqrt(phat*(1-phat)/1000)

% There is approximately a 39% chance that the (log) number of airl
% ine passengers will meet or exceed 7 in the next 5 years. The Monte
% Carlo standard error of the estimate is about 0.02
phat =
   0.3950
err =
   0.0155
```

11.5.2 绘制未来乘客的分布

根据仿真结果绘制未来 60 个月的乘客数（对数）分布。

```
figure
hist(Ysim(60,:))
title('60 个月的乘客数分布')
```

运行本节程序，会得到如图 11-10 所示的结果分布，乘客数基本符合正态分布。

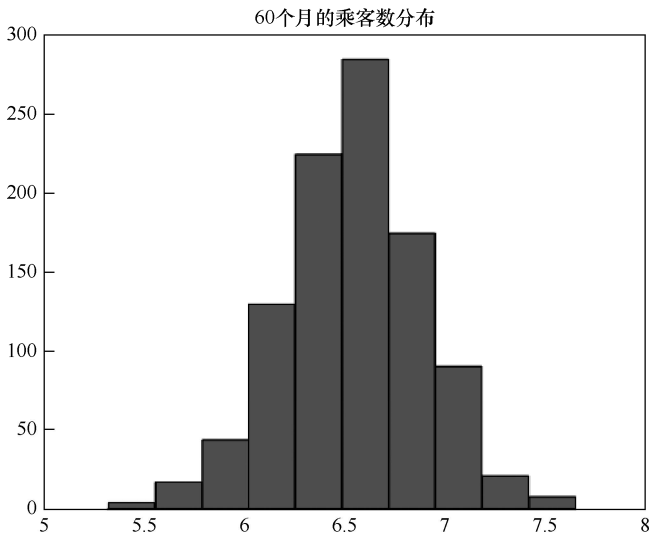


图 11-10　预测结果的分布

11.5.3　比较 MMSE 和蒙特卡罗仿真

在相同的预测时间范围内进行 500 次仿真，将仿真结果与 MMSE 预测进行比较。

```
rng('default')
res = infer(fit,Y);
Ysim = simulate(fit,60,'numPaths',500,'Y0',Y,'E0',res);

Ybar = mean(Ysim,2);
simU = prctile(Ysim,97.5,2);
simL = prctile(Ysim,2.5,2);

figure
h1=plot(Yf,'Color',[.85,.85,.85],'LineWidth',5);
hold on
h2 = plot(Ybar,'k--','LineWidth',1.5);
xlim([0,60])
plot([upper,lower],'Color',[.85,.85,.85],'LineWidth',5)
plot([simU,simL],'k--','LineWidth',1.5)

title('MMSE 和蒙特卡罗预测的比较')
```

```
legend([h1,h2],'MMSE','蒙特卡罗','Location','NorthWest')
hold off
```

MMSE 预测和模拟平均值实际上是难以区分的，如图 11-11 所示，理论 95%的预测区间和基于模拟的 95%的预测区间之间存在细微的差异。

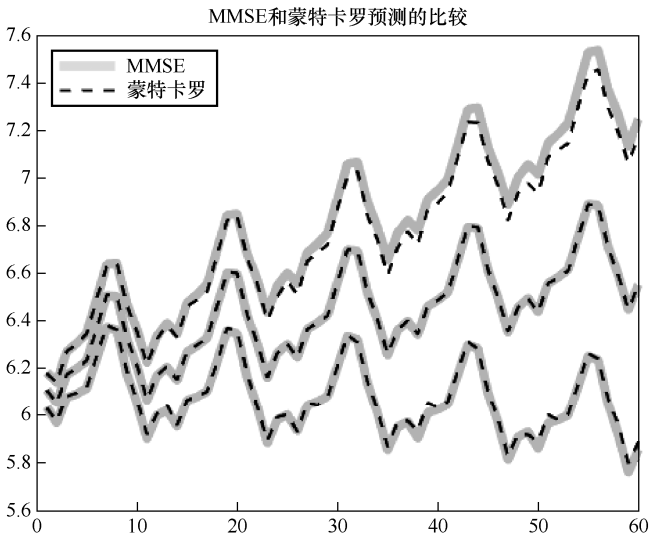


图 11-11　MMSE 与蒙特卡罗预测的比较

11.6　小结

本章介绍了用 ARIMA 模型对时间序列进行预测的全过程，包括：对样本进行分析，探索时序的基本特征，指定模型，估计模型参数，对模型进行测试，使用模型进行预测，对模型进行评估。使用 MATLAB 中的时间序列函数可以方便整个过程的实现，也便于进行数据的分析，还可以灵活地对数据进行其他形式的分析，如蒙特卡罗仿真。整个案例在时间序列的实际应用中非常具有典型性，值得借鉴。

12 股票收益时间序列的建模与预测

对于某些金融时间序列，当前的回报与以往的回报相关，这种行为可以使用自回归滑动平均模型 ARIMA 来捕获。另外，可以通过引入连续观测的差分来建模，以得到一个平稳序列。这些模型可以组合成 AR、MA、ARMA 或 ARIMA 模型。

金融时间序列也表现出波动聚类行为，即波动率高的时期与低波动时期。这种行为经常出现在每日或每周的数据中，而在月度数据中则较少出现。

引入广义自回归条件异方差 GARCH 模型的目的是获取财务收益这方面的信息，具体地说，通过允许条件方差随时间变化来捕获条件异方差。模型采用自回归形式，并且定义其方差依赖于过去时间序列的实际方差和之前评估的方差。这样的 GARCH 模型在表单中被参数化为（*i*,*j*），其中“*i*”表示模型中包含的滞后条件方差的数量，而“*j*”表示滞后信息的含义为 *t* 时刻的值与时间序列均值的差。

本章将通过一个具体的关于股票收益时间序列建模与预测的实例来介绍如何对金融指数利用 ARIMA 和 GARCH 两套模型来建模，并将使用模型和蒙特卡罗模拟对时间序列进行预测。

12.1 时序数据的获取与预处理

12.1.1 获取金融数据

使用 MATLAB Datafeed 工具箱，可以从联邦储备经济数据服务（FRED）中检索标普 500 指数的数据。Datafeed 工具箱支持从各种数据提供商获取财务数据，包括 Bloomberg、FactSet、ThomsonReuters、Yahoo Finance 和 Wind。

我们将获取 2001 年 1 月 1 日至 2013 年 12 月 31 日标普 500 指数的每日数据，并使用 MATLAB 来可视化从 FRED 服务器下载的数据。

```
clc, clear, close all
fromDate = '2001/01/01';
toDate = '2013/12/31';
Symbols = {'SP500'};
EconData = datafromFRED(fromDate, toDate, Symbols);
% Visualize data using custom visualization function
plotecondata(EconData,'Daily Data')
```

运行本节代码，可以下载到标普 500 的数据，并可以得到如图 12-1 所示的数据可视化结果。

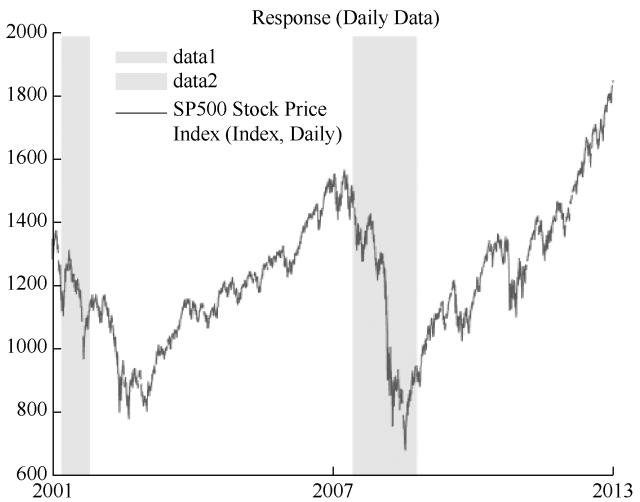


图 12-1　标普 500 走势图

12.1.2　数据的预处理

从 FRED 得到的标普 500 指数是以指数的形式出现的，可以将其转换为一个收益序列，并在数据不可用的时候过滤掉日期。

```
Dates = EconData.SP500.Data(:,1);
Index = EconData.SP500.Data(:,2);
Returns = tick2ret(Index);
% Filter out NaNs
```

```
idx = ~isnan(Returns);
Returns = Returns(idx); Dates = Dates(idx); Index = Index(idx);
% We plot the S&P 500 index and realized daily returns against time.
plotIndexReturns(Dates, Index, Returns)
```

运行本节代码，可以得到如图 12-2 所示的收益序列图。

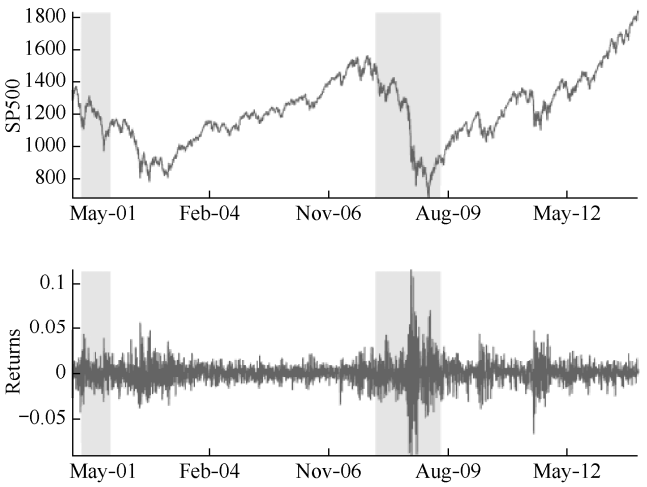


图 12-2　标普 500 收益序列图

12.2　时序数据分析

12.2.1　划分训练集和测试集

将数据集分成训练集和测试集，训练集用于确定模型参数，测试集用于测量模型的有效性。在本例中，训练集将使用前 60%的数据，而测试集将使用剩下 40%的数据。

```
trgFrac  = round(0.6 * length(Returns));
TrainingData = Returns(1:trgFrac); % First 60%
TestData = Returns(trgFrac+1:end); % Remaining 40%
plotIndexReturns(Dates, Index, Returns, trgFrac)
```

运行本节代码，可以得到训练集和测试集，如图 12-3 所示。

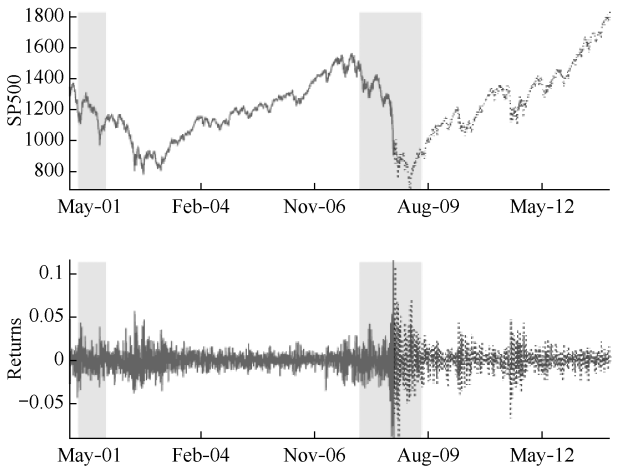


图 12-3　训练集和测试集的划分（右侧虚线部分为测试集）

12.2.2　测试序列的平稳性

一个时间序列被认为是静止的，如果它的某些或全部的时刻不是时间的函数，数据可以通过使用增广的 Dick-fuller 测试或 Phillips-Perron 测试来测试其平稳性。第二种测试更适用于残差异方差的数据，这在金融时间序列中经常出现。

“pptest”功能允许我们测试一个时间序列，以确定一个单位根的存在。当数据的自回归模型的特征方程在单位圆上有根时，单位根测试的原假设是：数据存在单位根，如果不拒绝，就可以表示一个单位根的存在。另外，“kpsstest”函数可以查找统计证据，以反对趋势平稳性的原假设，拒绝则意味着基础数据不是趋势平稳。

同时，两项测试的结果都表明，标准普尔 500 指数不是趋势平稳，而且可能存在一个单位根，这将通过数据的滞后表现出来。最后，使用“lmctest”函数在不同的索引数据（收益）上执行 Leybourne-McCabe 平稳性测试。具体地说，它检验数据是一个趋势平稳过程的假设，而不是数据是一个非平稳过程的假设。对于标准普尔 500 的数据，LMC 测试接受了另一种假设，即价格指数是非平稳的，然而，测试表明不同的指数可能是静止的。

用 MATLAB 实现这些测试和检验的脚本如下：

```
disp('Failing to reject the null could indicate the possible existe-
     nce of a unit root in the index.')
[hPP, pPP] = pptest(Index, 'model', 'TS') %#ok<*NOPTS>
disp('Rejection of the null implies the the index is not trend
     stationary.')
[hKPSS, pKPSS] = kpsstest(Index, 'trend', true)

disp(['The test rejects the null hypothesis for the differenced
data,'...
    'accepting the alternative that the differenced data might be
   stationary.'])
hPP_diff = pptest(diff(Index))

disp('LMC test accepts the alternative hypothesis that the index
     is non-stationary.')
[hLMC, pLMC] = lmctest(Index)
disp('LMC test indicates that the differenced data might be station-
     ary.')
[hLMC_diff, pLMC_diff] = lmctest(diff(Index))
```

运行本节脚本，得到如下结果：

```
hPP =
  logical
   0
pPP =   0.6795
%Rejection of the null implies the index is not trend stationary.
hKPSS =
  logical
   1
pKPSS =    0.0100
%The test rejects the null hypothesis for the differenced data,ace-
%pting the alternative that the differenced data might be stationary.
hPP_diff =
  logical
   1
%LMC test accepts the alternative hypothesis that the index is non-
%stationary.
hLMC =
  logical
   1
```

```
pLMC =
    0.0100
LMC test indicates that the differenced data might be stationary.
hLMC_diff =
  logical
   0
pLMC_diff =
    0.1000
```

12.2.3 数据的自相关性

现在我们可以探究收益序列的自相关和部分自相关属性，具体实现代码如下。

```
figure;
subplot(2,1,1);
autocorr(TrainingData);
subplot(2,1,2);
parcorr(TrainingData);
```

运行本节脚本，得到如图 12-4 的 ACF 和 PACF 图，由这两幅图可以发现在滞后 1 和 2 上有显著的相关性。

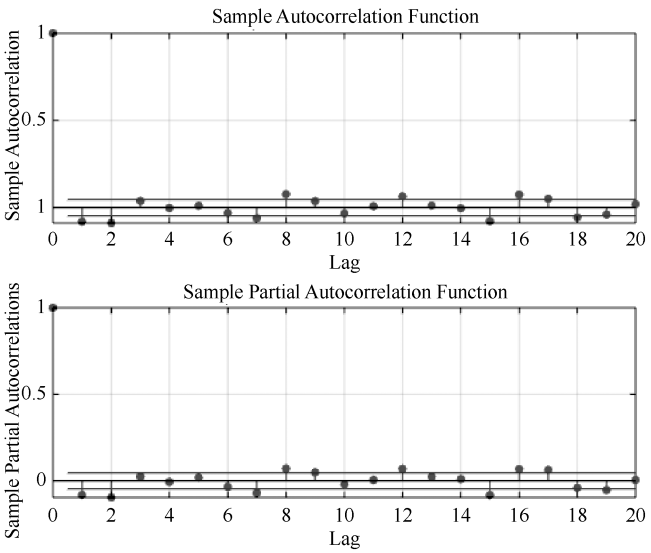


图 12-4　序列的 ACF 和 PACF 图

12.2.4 自相关性的统计测试

Ljung-Box Q 测试用于测试模型残差有没有自相关，原假设没有显著的自相关，具体实现代码如下。

```
adjTrainingData = TrainingData - mean(TrainingData);
disp(['The p-values obtained from the LBQ test support our initial '...
   'hypothesis of autocorrelations existing at lag orders 1 and 2.']);
[hLBQ, pLBQ] = lbqtest(adjTrainingData, 'Lags', 1:2, 'alpha', 0.05)
```

运行本节脚本，得到如下结果：

```
The p-values obtained from the LBQ test support our initial
hypothesis of autocorrelations existing at lag orders 1 and 2.
hLBQ =
  1×2 logical 数组
   1   1
pLBQ =
   1.0e-03 *
    0.4763    0.0015
```

12.3 模型估计

12.3.1 指定一个 ARIMA 模型

ARIMA 模型在计量经济学工具箱中可以使用“arima”和“estimate”函数来指定其参数。我们将尝试一个 ARMA(1,1)模型。arima 函数允许我们指定模型的参数，estimate 函数计算参数值，让模型更适合数据集，具体实现代码如下。

```
mdl = arima('ARLags', 1, 'MALags', 1);
[~, ~, logL, info] = estimate(mdl, TrainingData);
```

运行本节脚本，得到模型的具体参数：

```
ARIMA(1,0,1) Model (Gaussian Distribution):
           Value      StandardError    TStatistic      PValue
          _________    _____________    __________    _________
```

```
Constant     -4.45e-05     0.00012962      -0.34331       0.73136
AR{1}         0.47808       0.056718        8.4291     3.4825e-17
MA{1}        -0.56931       0.051486       -11.057     2.0174e-28
Variance   0.00014667     2.149e-06         68.25             0
```

12.3.2 估计 ARIMA 模型的滞后序列

不同滞后参数的 ARIMA 模型可以通过测量其信息准则系数来计算最优滞后参数，如 Akaike 信息准则或贝叶斯信息准则。在这里，为了计算标准普尔 500 的 ARMA 模型的最优滞后参数，我们通过改变 AR 滞后和 MA 从 1 到 8 的变化来执行参数扫描，并计算每对的 AIC 数字。

执行这些估计计算通常需要花费大量的时间。如果我们能够访问多核系统，就可以通过调用“parpool”命令并使用“parfor”而不是“for”来并行化“for”循环。这样做使计算性能得到提升，并能够最大限度地利用计算资源。

实现代码如下：

```
maxLags = 4;
AICSet = nan(maxLags, maxLags);

parfor i = 1:maxLags
    for j = 1:maxLags
        mdl = arima('ARLags',i,'MALags',j);
        [~, ~, logL, info]=estimate(mdl, TrainingData, 'display',
        'off');
        AICSet(i, j) = aicbic(logL, length(info.X));
    end
end
% We draw a heatmap that visually represents how AIC numbers vary.
% By visual inspection,we find that the minimum AIC number corresponds
  to
% an AR lag order of 2 and an MA lag order of 2. In other words,
% the best model to employ is an ARMA(2,2) model.
% figure;
```

```
    heatmap(AICSet/1000,1:size(AICSet,1),1:size(AICSet,2),'%0.2fK',
           'Gridlines',':','Colorbar',true);
    xlabel('MA Lags')
    ylabel('AR Lags')
    title('Akaike information criteria')
    [OptimalARLags, OptimalMALags] = find(AICSet==min(min(AICSet)));
    title(['Optimal AR and MA Lags are (' num2str(OptimalARLags) ','
    num2str(OptimalMALags) ')'])
```

运行本节脚本，可以得到如图 12-5 所示的参数扫描结果。

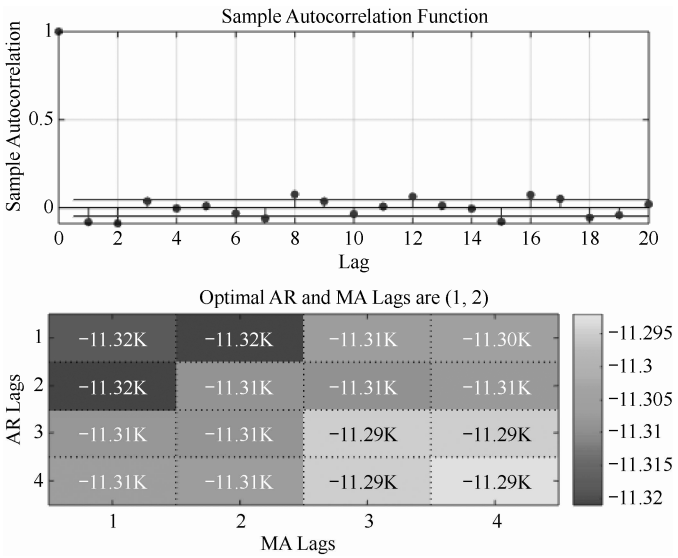


图 12-5　MA 与 ARL 参数扫描结果

12.4　模型的测试

12.4.1　测量残差的异方差性

到目前为止，模型没有考虑到方差的变化，也就是说，我们假设了同方差的条件。同时，异方差也是一种收益序列的行为，该序列的子部分表现出与整个系列不同的方差。

12.4.2 残差异方差的统计测试

统计测试，如 Engle 的 ARCH 测试，可以用来检验残差的异方差性，通过估计最合适的 ARMA 模型来计算残差。原假设是数据中没有条件异方差。具体实现代码如下。

```
mdl = arima('ARLags', OptimalARLags, 'MALags', OptimalMALags);
fit = estimate(mdl, TrainingData);
[res, ~, ~] = infer(fit, TrainingData);

disp('Rejection of the null implies that ARCH effect may exist in
     the data.')
[h, pVal] = archtest(res)
```

运行本节脚本可得：

```
ARIMA(1,0,2) Model (Gaussian Distribution):
             Value      StandardError    TStatistic     PValue
           __________    _____________    __________    __________

Constant   -8.0235e-05    0.00026238      -0.3058       0.75976
AR{1}       -0.086504       0.01517      -5.7024      1.1816e-08
MA{2}        -0.094493     0.0098379        -9.605     7.6163e-22
Variance     0.00014644    2.2422e-06       65.309              0
h =
  logical
   1
pVal =
   1.6488e-10
```

12.5 GARCH 模型的估计

12.5.1 估计 GARCH 的滞后序列

GARCH 模型可以被参数化为（*p*,*q*），其中“*p*”表示 garch 模型条件方差的滞后

阶数，q 表示条件异方差模型 ARCH 的滞后阶数。

我们将尝试估计一个组合的条件均值和条件方差模型：模型的 ARMA 部分的参数之前估计的为（2,2）。在 ARMA 参数估计过程中，将对 ARCH 滞后和 GARCH 在 1～3 进行参数扫描。具体实现代码如下。

```
Recompute = false;

maxLags = 3;
AICSet = nan(maxLags, maxLags); % Preallocate fit statistics

if Recompute == true
   tic
   parfor p = 1:maxLags
      for q = 1:maxLags
         mdl = arima('ARLags', OptimalARLags, 'MALags', Optimal-
         MALags, ...'Variance', garch(p,q));
        [~,~,logL,info] = estimate(mdl, TrainingData, 'display',
        'off');
         AICSet(p,q) = aicbic(logL, length(info.X));
      end
   end
   toc
   save garchAIC AICSet
else
   load garchAIC
end

figure;
heatmap(AICSet/1000,1:size(AICSet,1),1:size(AICSet,2),'%0.2fK',
'Gridlines',':','Colorbar',true)
xlabel('ARCH Lags')
ylabel('GARCH Lags')
title('Akaike information criteria')
```

```
[garchLags, archLags] = find(AICSet==min(min(AICSet)));
title(['Optimal GARCH and ARCH Lags are (' num2str(garchLags) ','
num2str(archLags) ')'])
```

运行本节代码可得如图 12-6 所示的参数扫描结果。

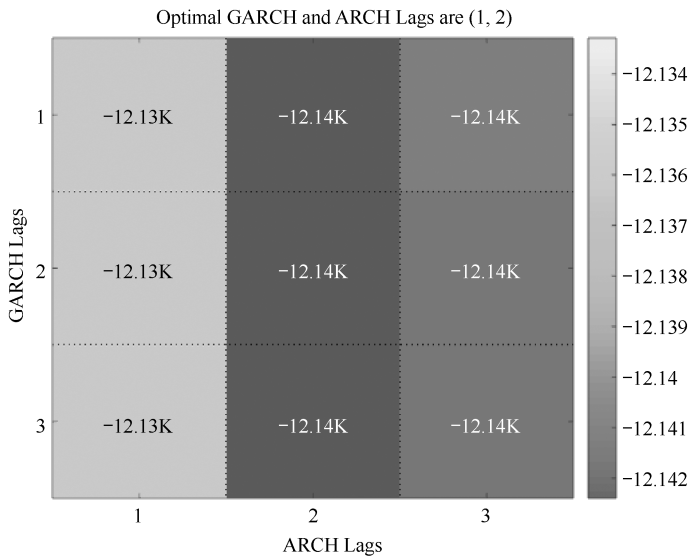


图 12-6　ARCH Lags 与 Garch Lags 参数扫描结果

12.5.2　组合条件 ARMA/GARCH 模型

现在可以定义一个组合的模型并设置最优参数。

```
mdl = arima('ARLags', OptimalARLags, 'MALags', OptimalMALags, ...
          'Variance', garch(garchLags, archLags));
fit = estimate(mdl, TrainingData);
```

运行如下脚本：

```
ARIMA(1,0,2) Model (Gaussian Distribution):

              Value       StandardError    TStatistic     PValue
            ___________   _____________    ___________   _________

Constant    0.00034104     0.00019446        1.7538       0.079472
```

```
AR{1}       -0.067805      0.023355      -2.9032      0.0036938
MA{2}         -0.0594      0.024534      -2.4212       0.01547

GARCH(1,2) Conditional Variance Model (Gaussian Distribution):

              Value       StandardError    TStatistic      PValue
            ___________    _____________    ___________    ___________

Constant    1.4622e-06     6.7127e-07        2.1782        0.02939
GARCH{1}       0.89319        0.01329        67.207              0
ARCH{1}      0.0033353       0.013477       0.24747        0.80454
ARCH{2}       0.094125       0.017424         5.402      6.5902e-08
```

12.6 模型的仿真

12.6.1 收益、条件方差样本路径的仿真

可以使用组合条件均值和条件方差模型来产生样本，从而仿真出多个时序未来的变化路径，而已观测到的收益、残差和条件方差可以作为仿真的初值。实现代码如下：

```
rng('default');
periods = 252*5;
paths = 1000;
[E0, V0] = infer(fit, TrainingData);
[Y, E, V] = simulate(fit, periods, 'NumPaths', paths, ...
                  'Y0', TrainingData, 'E0', E0, 'V0', V0);

% We can use the computed simulation paths to plot a 95% confidence
% interval for forecasted returns beyond the training data.

% Plot simulated returns at 2.5, 50 and 97.5 percentile of all the paths
plotSimData(Dates, TrainingData, prctile(Y,[2.5 50 97.5],2))
ylabel('Returns')
```

```
% We can also plot simulated conditional variances over the same
%horizon
plotSimData(Dates, V0, prctile(V,[2.5 50 97.5],2))
ylabel('Conditional Variances')
```

运行本节脚本可得如图 12-7 所示的收益仿真结果，以及如图 12-8 所示的条件方差仿真结果。

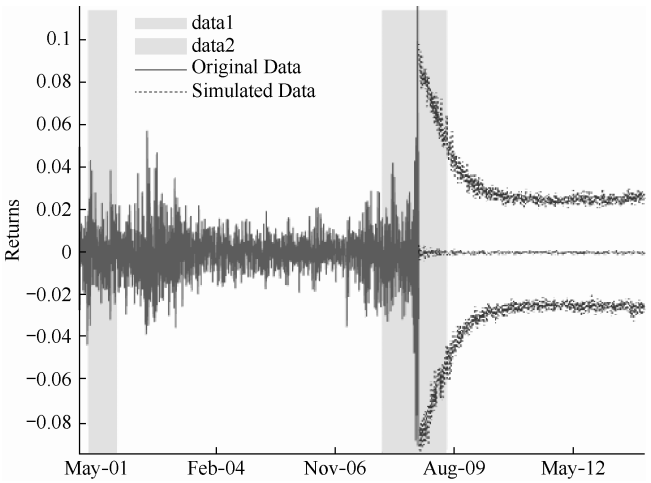


图 12-7　收益的仿真结果

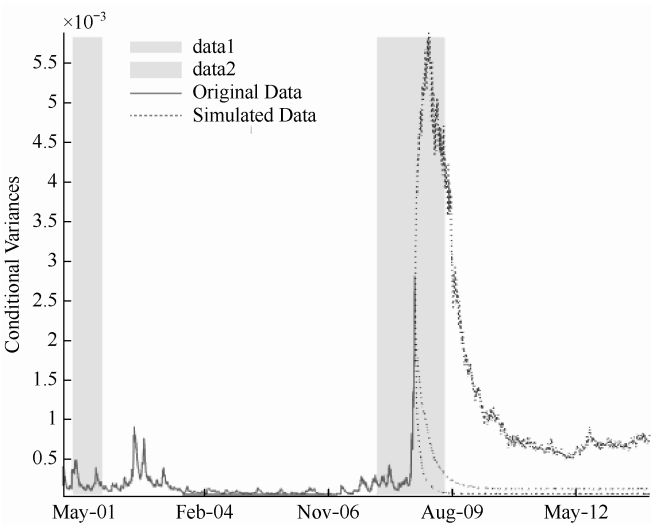


图 12-8　条件方差仿真结果

12.6.2 蒙特卡罗预测

我们还可以使用滚动窗口进行基于蒙特卡罗预测，通过使用“模拟”函数来计算多个样本路径。仿真平均值为基础预测，仿真实现的 2.5、97.5 百分位数可以作为近似 95%预测区间的边界路径。

在这里，我们提供了一个长达 252 天的滚动窗口（numDataPoints），每次向前移动 1 天。我们执行蒙特卡罗模拟来生成 1000 条随机路径，并选择 2.5、50 和 97.5 百分位数。实际上，我们从培训数据的末尾生成了一个 Numtimeday 预测，并可以将该预测与测试数据进行比较。

完整的模拟需要一些时间来运行。因此，我们保存了完整运行的结果，并可以使用这些预测。可以将布尔标志“usePrecomputed”设置为“true”，以便使用预先计算的数据。使用并行计算工具箱来加速这些计算，因为平均速度几乎与可用的 MATLAB 计算核的数量成线性关系。实现代码如下。

```
Recompute = false;

% Simulation parameters
numTimeSteps = 1; % Simulate one day at a time and re-estimate the model
numPeriods = 252*5; % Simulate five years into the future
paths = 1000;
numDataPoints = 252*1; % Use last year of data to estimate the model

Ysimulations = nan(numPeriods,paths);
ret = slicereturns(numDataPoints,numPeriods,Returns,length(Tra-
iningData)); % For speeding up PARFOR
rng default % for repeatability

if Recompute == true
   tic
   parfor i = 1:numPeriods
      fit = estimate(mdl, ret(:,i), 'display', 'off');
      [E0, V0] = infer(fit, ret(:,i));
      Ysimulations(i,:) = simulate(fit, numTimeSteps, 'NumPaths',
      paths,...'Y0', ret(:,i), 'E0', E0, 'V0', V0);
```

```
    end
    toc
    save PrecomputedSimulation Ysimulations
else
    load PrecomputedSimulation
end

% We can use the computed simulation paths to plot a 95% confidence
% interval for forecasted returns beyond the training data.

% Plot simulated returns at 2.5, 50 and 97.5 percentile of all the paths
plotSimData(Dates, TrainingData, prctile(Ysimulations,[2.5 50
            97.5],2))
ylabel('Returns')
```

运行本节脚本可得如图 12-9 所示的蒙特卡罗仿真结果。

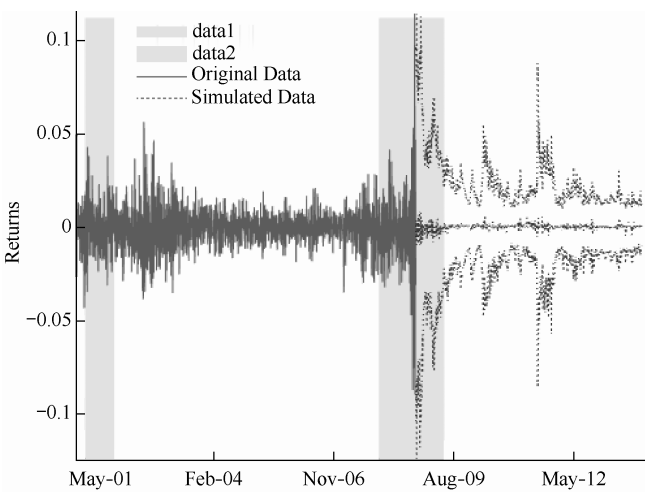


图 12-9　蒙特卡罗仿真结果

12.6.3　可视化价格轨迹

返回的计算模拟路径可用于为每条路径生成价格预测，实现代码如下：

```
ProjectedPrice = ret2tick(Ysimulations,Index(length(TrainingData)));
```

```
% Plot simulated data
plotSimData(Dates, Index(1:trgFrac), ProjectedPrice(2:end,:))
% Plot original data next to simulated data for visual comparison
plot(Dates(trgFrac+1:trgFrac+numPeriods),Index(trgFrac+1:trgFra
     c+numPeriods))
ylabel('S&P 500')
title('Realized vs All Forecasted Paths');
```

运行本节脚本可得如图 12-10 所示的预测轨迹分布图。

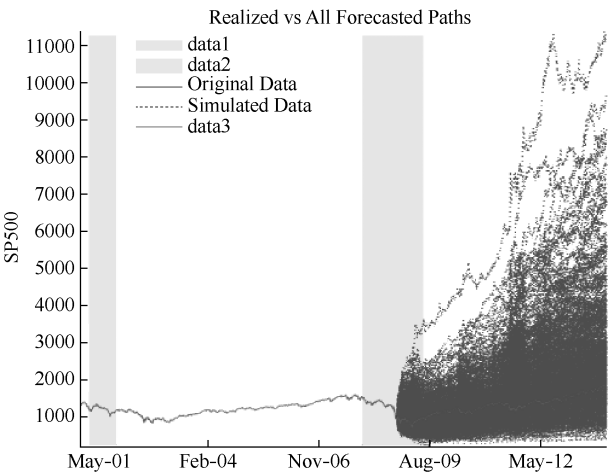


图 12-10　预测轨迹分布图

12.6.4　可视化预测中值路径

我们只考虑预测的中值价格路径，结果表明：在不那么极端的情况下，它会更平稳。可以修改 percentile 参数“prctileParam”来研究其他价格路径。实现代码如下。

```
prctileParam = 50; % For Median Price
mProjectedPrice = prctile(ProjectedPrice,prctileParam,2);
% Median Price

% Plot simulated data
plotSimData(Dates, Index(1:trgFrac), mProjectedPrice(2:end,:))
% Plot original data next to simulated data for visual comparison
```

```
    plot(Dates(trgFrac+1:trgFrac+numPeriods),Index(trgFrac+1:trgFra
    c+numPeriods))
    ylabel('S&P 500')
    title('Realized vs Median Forecasted Path');
```

运行本节脚本可得如图 12-11 所示的预测结果。

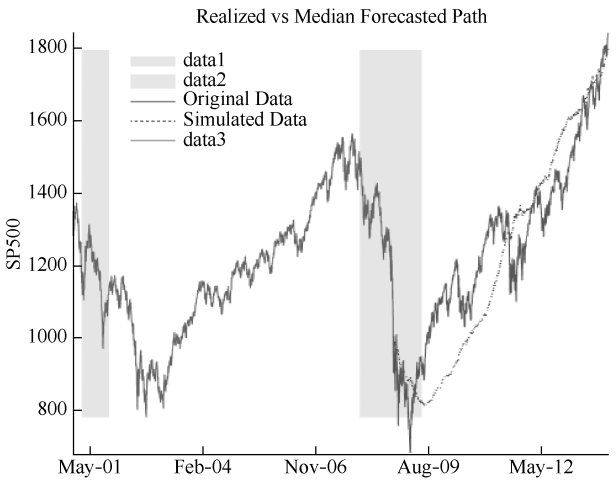


图 12-11　中值价格路径预测结果

12.7　小结

本章通过一个具体的对金融时间序列建模的实例，介绍了如何针对实际问题综合应用 ARIMA 和 GARCH 两套模型取得更好的应用效果。从这个案例来看，两套模型都有自己的特点，但并不是完全独立的，对一个问题同时使用两套模型，可以辨识出数据的不同特征，对发现数据特征及更好地进行建模和预测是非常有帮助的。